UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

Estudo da vinculação de um Cliente Particular a um Banco

Helena Maria Alves de Almeida Carvalho

Mestrado em Matemática Aplicada à Economia e Gestão

Trabalho de Projeto orientado por:
Raquel João Fonseca

2017

Dedicatória e agradecimentos

Dedico este trabalho aos meus pais e ao meu namorado por todo o apoio e motivação que ao longo destes anos demonstraram ter para comigo. À minha mãe por me abrir os horizontes, por ser a minha melhor amiga e conselheira, pela compreensão e pela disciplina que sempre incutiu em mim. Ao meu pai por ter sido o primeiro a ensinar-me o gosto pela Matemática, por ter sido o meu primeiro professor, pelo incentivo, pela proteção e por saber que eu conseguia dar mais do que aquilo que fui capaz de dar. Ao António por ser o meu porto de abrigo e a minha fonte de confiança, por ter sempre uma palavra a dizer e por valorizar tudo aquilo a que me dediquei.

Agradeço a Deus por me guiar e iluminar ao longo desta caminhada e por não me ter deixado desanimar perante as dificuldades.

Aos meus tios e padrinhos agradeço por mesmo estando longe nunca faltarem com o seu apoio e suporte familiar.

Agradeço à minha professora e orientadora, Doutora Raquel Fonseca, que sempre vi como um exemplo a seguir, por toda a amabilidade e disponibilidade que demonstrou desde o início e por todas as críticas e conselhos que simplificaram os problemas que surgiram.

Agradeço à professora Doutora Teresa Alpuim por toda a aprendizagem que me permitiu alcançar, pela exigência e por estar sempre próxima dos seus alunos. E a todos os professores que acompanharam o meu percurso académico agradeço todos os ensinamentos e princípios transmitidos.

Ao Dr. Pedro Dinis e a toda a equipa com quem trabalhei agradeço por me terem permitido desenvolver este projeto e por terem sido as primeiras pessoas que no mercado de trabalho puxaram por mim, ensinaram e acompanharam no conhecimento de novos desafios e de novas experiências. Agradeço ainda à Paula Silva por ter sido minha "mentora" e por me ter ensinado novos conceitos e técnicas para o desenvolvimento deste projeto.

Por último e por nada disto ser possível se estivesse sozinha agradeço aos pais do António por estarem sempre presentes e por todo o apoio. Não podendo faltar o agradecimento à minha segunda família, aquela que escolhi, que são os meus amigos. Não menciono o nome mas eles sabem o quão importantes são. Estiveram sempre ao meu lado ao longo destes anos. Obrigada pela força, pela amizade, por acreditarem e por terem proporcionado todas as alegrias e sorrisos que animaram os meus dias.

"*Cada um que passa na nossa vida passa sozinho, pois cada pessoa é única, e nenhuma substitui outra. Cada um que passa na nossa vida passa sozinho, mas não vai só, nem nos deixa sós. Leva um pouco de nós mesmos, deixa um pouco de si mesmo. Há os que levam muito; mas não há os que não levam nada. Há os que deixam muito; mas não há os que não deixam nada. Esta é a maior responsabilidade de nossa vida e a prova evidente que duas almas não se encontram ao acaso.*"

Saint-Exupéry

Resumo

Através de um estágio no Banco B na área de Data Mining e Estudos de Mercado, este trabalho foi feito para a obtenção do grau de mestre em Matemática Aplicada à Economia e Gestão, pela Faculdade de Ciências da Universidade de Lisboa.

O presente trabalho consiste na realização de um Modelo de Regressão Logística com a finalidade de estudar a vinculação de um Cliente a um banco. Foi feito um estudo de mercado onde o principal objetivo é saber o perfil de cada Cliente que considera um determinado banco como o seu banco principal. O Modelo de Regressão Logística terá como variável resposta a vinculação, ou não, de um Cliente a um banco consoante as variáveis que melhor caracterizam os seus perfis.

No Modelo Logístico existe a particularidade de, através de um conjunto de variáveis independentes, se conseguir prever a vinculação de um determinado Cliente. Esta é a melhor metodologia adotada para este estudo uma vez que interpreta o impacto marginal de cada variável na vinculação do Cliente.

Como tal este trabalho é composto por 4 capítulos. No primeiro é feita a introdução aos conteúdos teóricos da regressão logística múltipla. No capítulo seguinte será explicado o objetivo deste estudo e uma análise descritiva dos dados. Em seguida será posto em prática o Modelo de Regressão Logística através do uso de dados reais e de técnicas como a discretização de variáveis, árvores de decisão, R-Quadrado, Qui-Quadrado e o Stepwise (método que seleciona as variáveis finais do modelo). Neste último capítulo é feita ainda uma avaliação do modelo e consequente análise do perfil de Clientes e validação do modelo construído.

Por fim será apresentado o capítulo que conclui e comenta o trabalho feito ao longo do modelo e ainda sugere algumas alternativas metodológicas tais como Splines Cúbicas e ainda um modelo alternativo.

Palavras-chave: Modelo de Regressão Logística Múltipla, Discretização, Stepwise, Estudos de Mercado, Data Mining

Abstract

Through an internship at Bank B in the area of Data Mining and Market Research, this work was done to obtain a master's degree in Mathematics Applied to Economics and Management, by the Faculty of Sciences of the University of Lisbon.

The present work focuses on the utilization of a Logistic Regression Model with the purpose of studying the connection of a Client to a bank. A market research was done where the main objective is to know the profile of each Client that considers a bank as his/her main bank. The Logistic Regression Model will have as response variable the binding, or not, of a Client to a bank according to the variables that best characterize his/her profile.

In the Logistic Model, there is the particularity of, through a set of independent variables, predicting the binding of a certain Client. This is the best methodology adopted for this study since it interprets the marginal impact of each variable in the Customer's binding.

As such this work is composed of 4 chapters. In the first chapter, the theoretical contents of Multiple Logistic Regression are introduced. In the following chapter, we will explain the purpose of this study and a descriptive analysis of the data. Then, in the third chapter, the Logistic Regression Model will be implemented using real data and techniques such as discretization of variables, decision trees, R-Square, Chi-Square and Stepwise (method that selects the final variables of the model). In this last chapter is done an evaluation of the model and consequently analysis of the profile of each Client and validation of the built model.

Finally, in the fourth, we will present the chapter that concludes and comments all the work done throughout the model and suggests some methodological alternatives such as Cubic Splines and an alternative model.

Keywords: Multiple Logistic Regression Model, Discritization, Stepwise, Market Research, Data Mining

Índice

Lista de figuras, gráficos e tabelas

Figuras

Gráficos

Tabelas

Introdução

A banca é um setor em permanente mudança, onde o Cliente assume um papel central. Agradar e satisfazer os Clientes, prestando um serviço de excelência e apresentando propostas de valor atrativas (inclusive com vista à criação de valor para o Banco) é uma preocupação constante dos bancos, que hoje em dia, apresentam planos de expansão ambiciosos.

Para tal existem áreas que elaboram estudos de mercado dedicados ao Cliente, muitas vezes ligados ao Departamento de Marketing. Cada vez mais o Marketing desempenha um papel importante por suportar toda a estratégia, a técnica e os mecanismos que regem as relações de troca (bens, serviços ou ideias) e pretende que o resultado de uma relação seja uma transação satisfatória para todas as partes que participam no processo. O Marketing comporta-se como uma atividade a médio e longo prazo que finda assegurar a obtenção do maior benefício possível pretendendo maximizar o consumo, a satisfação do consumidor, a escolha e a qualidade de vida. É neste contexto que são aplicados conhecimentos avançados a respeito da prospeção de mercados e a sondagem de opiniões.

A fidelização e as relações de longo prazo com Clientes eram conceitos secundarizados não só pelas empresas como também pela comunidade científica. O marketing relacional foi abordado por Berry (1983), ainda no início da década de 80, como a ação de atrair, manter e melhorar as relações com os Clientes. Só nos finais das décadas de 80 e 90 é que o marketing relacional passou a exprimir-se não só no mundo científico como também no mundo empresarial. Segundo Berry, marketing de relacionamento é então a atração, manutenção e a potencialização de relações com os consumidores.

Para o marketing relacional será desejável que as empresas criem canais de relacionamentos com os seus Clientes, de modo a serem capazes de maximizar o seu valor. Para tal, deve ser realizada uma análise detalhada do mercado.

Define-se Estudo de Mercado como um processo sistemático e objetivo de recolha e consequente fornecimento da informação necessária e indispensável para a tomada de decisões por parte da gestão/direção de marketing (Lopes, 2007). Os Estudos de Mercado auxiliam o marketing a detetar e avaliar qualitativa e quantitativamente as necessidades e/ou preferências dos consumidores, bem como o impacto das ações de marketing levadas a cabo.

Por outro lado, para analisar e aprofundar toda a informação sobre cada Cliente é preciso olhar para uma grande base de dados e analisar os padrões de informação desconhecidos e inesperados. É o processo de extração de informações previamente desconhecidas, compreensíveis e acionáveis, de grandes bases de dados, que implica a tomada de decisões cruciais de negócio, que define em parte o Data Mining. Segundo Fayyad (1996), é um processo não trivial de identificação de padrões válidos, novos e potencialmente úteis. É uma área de pesquisa multidisciplinar que inclui inteligência artificial, estatística, reconhecimento de padrões, recuperação de informação, computação de alto desempenho e visualização de dados, entre outras, algumas delas utilizadas em estudos de mercado.

Assim, o concílio entre Data Mining e Estudos de Mercado é muito importante. Enquanto que o primeiro explora e relaciona a informação que o Cliente tem com a sua entidade, o segundo tem em atenção as atitudes e opiniões que um Cliente tem com o seu fornecedor e desta maneira consegue-se obter um estudo mais pormenorizado e detalhado sobre o perfil de cada Cliente.

Vinculação é um novo conceito numa empresa e ainda uma evolução no marketing relacional, da manutenção e retenção de Clientes (Vilares e Coelho, 2011). Supõe a união da satisfação do Cliente com uma ação de consumo estável e duradoura. Reflete-se na compra sistemática de aquisição de bens ou serviços por parte dos consumidores. As empresas preocupam-se com os Clientes estarem satisfeitos, assim como os Clientes vão acabar por recompensar as empresas continuando a preferi-las e a adquirir

os seus bens e serviços. Vinculação é o ato de tornar os Clientes em pessoas ligadas ao seu produto, marca ou serviço.

Para poder haver uma relação de vinculação é preciso que se consigam estabelecer laços e manter relacionamentos a longo prazo com os Clientes, principalmente com aqueles que são mais rentáveis e fiéis ao Banco. A satisfação do Cliente é um grande pilar, está no centro da atividade económica. Muitas vezes, satisfação e vinculação são conceitos que podem ser confundidos. Mas como já foi referido, vinculação é um relacionamento de longo prazo enquanto que a satisfação pode depender de uma única transação e das expectativas criadas sobre um produto ou serviço.

Do ponto de vista empresarial podem ser destacadas algumas vantagens, como a facilidade e a incrementação das vendas, a redução dos custos de promoção, a retenção de empregados, a menor sensibilidade ao preço e a prescrição pela parte dos consumidores vinculados. Do ponto de vista dos consumidores há uma redução no risco percebido, um serviço personalizado e evitam-se custos de mudança. O objetivo é que o próprio Cliente, para além de ser fiel e vinculado à empresa, se sinta unido e com um verdadeiro compromisso para com ela. Um Cliente vinculado garante à empresa uma próxima venda e é ainda um Cliente satisfeito.

Capítulo I

Neste capítulo inicializa-se a introdução aos Modelos Lineares Generalizados e, por sua vez, a um dos seus casos particulares, a Regressão Logística. Como o caso em estudo é feito através de um Modelo de Regressão Logística é necessário explicar as etapas envolventes: o ajustamento do modelo, seleção de variáveis, a interpretação dos coeficientes estimados e análise e avaliação do modelo construído.

Modelos Lineares Generalizados

A Regressão é um método que permite estudar e explorar a relação entre uma ou mais variáveis explicativas (variáveis independentes) e uma variável resposta (variável dependente). Assim surgem os modelos de regressão que constituem um importante papel na análise estatística de dados para se poder modelar relações entre variáveis. As relações entre a variável resposta e as covariáveis são feitas através de uma função linear de alguns parâmetros. Modelos que dependem de forma linear dos seus parâmetros desconhecidos são mais fáceis de ajustar e as propriedades estatísticas dos estimadores resultantes são fáceis de determinar.

Quando o estudo é entre uma variável explicativa e uma variável resposta trata-se de uma Regressão Linear Simples. Mas se for o caso de estudar a relação da variável resposta com várias variáveis independentes então utiliza-se a Regressão Linear Múltipla.

O modelo linear normal, "criado" no início do século XIX por Legendre e Gauss, dominou a modelação estatística até meados do século XX, embora vários modelos não lineares e não normais tenham entretanto sido desenvolvidos para fazer face a situações que não eram adequadamente explicadas pelo modelo linear normal. Exemplo disso, tal como referem McCullagh and Nelder (1989) e Lindsey (1997): o modelo *complementar log-log* para ensaios de diluição (Fisher, 1992), os modelos *probit* (Bliss, 1935) e *logit* (Berkson, 1994; Dyke and Patterson, 1952; Rash, 1960) para proporções, os modelos *log-lineares* para dados de contagens (Birch, 1963), e os modelos de regressão para análise de sobrevivência (Feigl and Zelen, 1965; Zippin and Armitage, 1966; Glasser, 1967).

Todos os modelos anteriormente descritos apresentam uma estrutura de regressão linear e têm em comum o facto da variável resposta seguir uma distribuição dentro de uma família de distribuições com propriedades muito específicas: a *família exponencial*.

Nelder e Wedderburn (1972) introduziram os Modelos Lineares Generalizados que correspondem a uma síntese destes e de outros modelos, vindo assim unificar, tanto do ponto de vista teórico como conceptual, a teoria da modelação estatística até então desenvolvida. São casos particulares dos modelos lineares generalizados, doravante referidos como MLG, os seguintes modelos:

- Modelo de Regressão Linear Clássico;
- Modelo de Análise de Variância e Covariância;
- Modelo de Regressão Logística;
- Modelo de Regressão de Poisson;
- Modelos *log-lineares* para tabelas de contingência multidimensionais;
- Modelo *probit* para estudos de proporções, etc.

Estes modelos permitem, por um lado, que a variável resposta não tenha distribuição normal, e por outro lado, que estejam relacionados com a variável resposta através de uma função de ligação. Tanto a variável resposta como as covariáveis podem ser de natureza contínua ou discreta (Turkman, 2000).

Os modelos lineares generalizados são então uma extensão do modelo linear clássico

$$Y = X\beta + \varepsilon,$$

em que X é uma matriz de dimensão $n \times (p+1)$ de especificação do modelo correspondendo, geralmente, à matriz de covariáveis (sendo o primeiro vetor unitário) e estando associada ao vetor de parâmetros $\beta = (\beta_0, \beta_1, \dots, \beta_p)^T$. O vetor de erros aleatórios é dado por ε com distribuição $N(0, \sigma^2 I)$, sendo I a matriz identidade.

Os MLG pressupõem que a variável resposta tenha distribuição pertencente a uma família particular, a família exponencial e são caracterizados pela seguinte estrutura:

1. Componente aleatória
   Dado o vetor de covariáveis x_i as variáveis Y_i são (condicionalmente) independentes com distribuição pertencente à família exponencial, com $E(Y_i|X_i) = \mu_i$.
2. Componente estrutural ou sistemática
   O valor esperado μ_i está relacionado com o preditor linear $\eta_i = z_i^T \beta$ através da relação:

$$\mu_i = h(\eta_i) = h(z_i^T \beta)$$
$$\eta_i = g(\mu_i),$$

   tal que h é uma função monótona e diferenciável;
   $g = h^{-1}$ é a função de ligação;
   β é um vetor de parâmetros de dimensão p;
   z_i é um vetor de especificação de dimensão p, função do vetor de covariáveis x_i.
   Geralmente $z_i = (1, x_{i1}, \dots, x_{ik})^T$ com $k = p - 1$.

Devido ao grande número de modelos que englobam e à facilidade de análise associada ao rápido desenvolvimento computacional que se tem verificado nas últimas décadas, os MLG têm vindo a desempenhar um papel cada vez mais importante na análise estatística, apesar das limitações ainda impostas, nomeadamente por manterem a estrutura de linearidade, pelo facto das distribuições se restringirem à família exponencial e por exigirem a independência das respostas. Há já atualmente, na literatura (Turkman, 2000), muitos desenvolvimentos da teoria da modelação estatística onde estes pressupostos são relaxados, mas o não acompanhamento dos modelos propostos com *software* adequado à sua fácil implementação, faz com que se anteveja ainda por algum tempo um domínio dos MLG nas aplicações de natureza prática.

Regressão Logística

Um dos casos particulares dos modelos lineares generalizados são então os modelos onde a variável resposta é uma variável binária (categórica), isto é, assume os valores "1" ou "0" se um determinado acontecimento se verifica ou não, respetivamente. Desses modelos, o modelo de Regressão Logística é o mais popular. Desde há duas décadas que o modelo de regressão logística se tornou, em muitos campos, o método de análise padrão. Foi através de Cox (1970) que passou a ser mais utilizada em estudos estatísticos.

Os modelos lineares assentam no pressuposto de que os termos de erro são variáveis independentes e identicamente distribuídas com distribuição normal. Consequentemente as observações da variável dependente são também independentes e com distribuição normal. Mesmo que os termos de erro se afastem da distribuição normal, para amostras de dimensão grande, utilizam-se os métodos de inferência estatística, não tendo o risco de cometer grandes erros. O facto dos modelos lineares terem então uma distribuição normal deve-se ao Teorema do Limite Central e suas generalizações para variáveis não identicamente distribuídas, e da Lei Fraca dos Grandes Números (Alpuim, 2014).

O objetivo da Regressão Logística é modelar, a partir de um conjunto de observações, a relação entre uma variável dicotómica e várias variáveis explicativas numéricas (contínuas e discretas) e/ou categóricas. É com um objetivo preditivo que o mercado tem utilizado cada vez mais esta técnica. Como o resultado da aplicação deste método reflete a probabilidade de ocorrência de um determinado evento, é feita uma predição no sentido de assumir que tal evento ocorrerá para os registos que possuam maior probabilidade e não ocorrerá para os demais.

Na Regressão Logística (Hosmer & Lemeshow, 1989), a variável resposta, Y, é uma variável com distribuição binomial de parâmetros 1 e π, ou seja, tem distribuição de Bernoulli, $Y \cap Ber(1,\pi)$, em que a função massa de probabilidade é dada por:

$$f(y|\pi) = \pi^{y}(1-\pi)^{1-y} \ , \qquad \text{para } y = 0{,}1$$

Da mesma maneira que o modelo de regressão linear pode ser ajustado para a variável resposta tendo em conta mais do que uma variável explicativa (covariável), na regressão logística também pode ser aplicada a mesma metodologia ao que se passa a chamar de Modelo de Regressão Logística Múltipla.

O vetor das variáveis independentes é representado por:

$$X^{T} = (X_1, X_2, \dots, X_p)$$

A probabilidade de ocorrência de um certo acontecimento varia consoante outras características associadas a esse acontecimento, ou seja,

$$\pi = \pi(X_1, X_2, \dots, X_p) = P(Y = 1|X_1, X_2, \dots, X_p)$$

Segundo Hosmer e Lemeshow (1989), em qualquer regressão a quantidade chave é o valor médio da variável resposta dado o valor da variável independente,

$$E(Y|X = x)$$

isto é, o valor médio condicional, onde Y representa a variável resposta e X a variável explicativa.

Assim sendo, é possível a média assumir qualquer valor quando x varia entre -∞ e +∞.

No modelo de Regressão Logística Múltipla, a probabilidade de se dar um determinado acontecimento pode ser escrita como uma função logística múltipla de um conjunto de variáveis independentes, que se escreve como:

$$\pi(X) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}$$

Neste caso, como a distribuição de Bernoulli pertence à família exponencial vem que

$$E(Y|X) = \pi$$

Com o objetivo de linearizar o modelo, uma das transformações fulcrais é a transformação $logit$, $g(x)$, ou seja, a função de ligação é dada por

$$g(x) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right)$$

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

O Modelo Linear Generalizado definido pelo modelo binomial, com função de ligação canónica logit é assim conhecido por Modelo de Regressão Logística (Turkman & Silva, 2000)

Para a construção do modelo logístico procede-se ao seu ajustamento onde é feita a estimação dos parâmetros através do método da máxima verosimilhança e posteriormente a uma seleção de variáveis que melhor explicam o modelo.

Ajustamento do modelo

Assumindo uma amostra de n observações independentes do par $(x_i, y_i),\ i = 1,2, \dots, n$ onde y_i é o valor da variável dicotómica e x_i o i-ésimo valor do vector de variáveis independentes, para ajustar o modelo é preciso estimar $\beta^T = (\beta_0, \beta_1, \dots, \beta_p)$, os parâmetros desconhecidos.

No caso múltiplo, o método utilizado para a estimação dos parâmetros é o Método da Máxima Verosimilhança tal como no caso univariado. A função de verosimilhança $L(\beta)$ representa a distribuição conjunta dos dados observados. Uma vez encontrada a função para um determinado conjunto de dados, o método da máxima verosimilhança determina a estimação de um conjunto de parâmetros desconhecidos que maximizam essa mesma função.

Para uma amostra de dimensão *n*, em que as observações são independentes, a função verosimilhança é dada por

$$L(\beta) = \prod_{i=1}^{n} f(y_i|\, x_i) = \prod_{i=1}^{n} \pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i}$$

Maximizar a função $L(\beta)$ é equivalente a maximizar o logaritmo de $L(\beta)$, o que simplifica muito os cálculos e, portanto, o logaritmo da função de verosimilhança (log-verosimilhança) é dado por:

$$\ln[L(\beta)] = \sum_{i=1}^{n} \{y_i \ln[\pi(x_i)]\,(1 - y_i)\ln[1 - \pi(x_i)]\} =$$

$$= \sum_{i=1}^{n} \left[y_i\beta_0 + y_i\beta_1 x_{i1} + \cdots + y_i\beta_p x_{ip} - \ln(1 + e^{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}})\right]$$

Admitindo que se verificam certas condições de regularidade (Sen & Singer, 1993) os estimadores de máxima verosimilhança para β são obtidos como solução do sistema das p equações normais

$$\frac{\partial ln L(\beta)}{\partial \beta_0} = \sum_{i=1}^{n} [y_i - \pi(x_i)] = 0$$

$$\frac{\partial ln L(\beta)}{\partial \beta_k} = \sum_{i=1}^{n} x_{ik}[y_i - \pi(x_i)] = 0$$

Em que $k = 1, \dots, p$, ou em notação matricial: $X'(Y - \Pi) = 0$.

Os estimadores de máxima verosimilhança de β são obtidos como solução das equações de verosimilhança. A solução não corresponde necessariamente a um máximo global da função $L(\beta)$. Contudo, em muitos modelos a função log-verosimilhança $L(\beta)$ é côncava, de modo que os máximos local e global coincidem. Para funções estritamente côncavas os estimadores de máxima verosimilhança são mesmo únicos quando existem. O problema da existência e unicidade destes estimadores ainda não tem uma teoria geral porque nem todos os modelos têm propriedades comuns no que diz respeito a esta questão. Partindo do princípio de que existe solução e que ela é única, subsiste ainda o problema com o cálculo das estimativas de máxima verosimilhança, pois as equações de verosimilhança não têm, em geral, solução analítica, implicando o recurso a métodos numéricos (Turkman, 2000).

No modelo de regressão linear as equações de verosimilhança são facilmente resolvidas. Para o modelo de regressão logística, tais equações são não-lineares nos parâmetros e desta forma, requer-se o uso de um procedimento iterativo. (Modelos de Regressão Logística por Cleonis Viater Fiqueira, Porto Alegre, 31 Março 2006).

Geralmente, estes estimadores têm distribuição conjunta assintoticamente normal e com matriz de covariâncias igual à inversa da matriz de informação de Fisher I(β) (ou matriz de covariância da função score), ou seja, são assintoticamente de variância mínima (Hosmer & Lemeshow, 2000).

A matriz de covariância dos coeficientes estimados é obtida a partir das derivadas parciais de segunda ordem do logaritmo da função verosimilhança:

$$\frac{\partial^2 \ln[L(\beta)]}{\partial \beta_j^2} = -\sum_{i=1}^{n}\left[x_{ij}^2\ \pi_i\ (1-\pi_i)\right]$$

$$\frac{\partial^2 \ln[L(\beta)]}{\partial \beta_j \beta_k} = -\sum_{i=1}^{n}\left[x_{ij}x_{ik}\pi_i(1-\pi_i)\right]\ ,$$

onde $j, k = 0,1,\dots,n$ e $\pi_i = \pi(x_i)$.

Desta forma a matriz I(β) é constituída pelo simétrico dos valores médios dos termos dados pelas equações anteriores. As variâncias e covariâncias entre os coeficientes estimados são obtidas através da inversa da matriz I(β), ou seja, $VAR(\beta) = I^{-1}(\beta)$. Consequentemente $VAR(\beta_j)$ designa o j-ésimo elemento da diagonal principal da matriz que é a variância de $\widehat{\beta_j}$, e $COV(\beta_j, \beta_w)$ são os elementos fora da diagonal da matriz que são a covariância entre $\widehat{\beta_j}$ e $\widehat{\beta_w}$. Os estimadores da variância e da covariância são obtidos de $VAR(\beta)$ quando se substitui β pelo seu estimador $\hat{\beta}$. Para designar os valores da respectiva matriz tem-se $\widehat{VAR}(\widehat{\beta_j})$ e $\widehat{COV}(\widehat{\beta_j}, \widehat{\beta_w})$, com $j, w = 0,1,2,\dots,p$.

O erro estimado dos coeficientes encontrados é representado por: $\widehat{SE}(\widehat{\beta_j}) = \sqrt{\widehat{VAR}(\widehat{\beta_j})}$.

Matricialmente vem que $\hat{I}(\hat{\beta}) = X'VX$, onde X é uma matriz $n \times (p+1)$ e V é uma matriz diagonal $(n \times n)$ de elemento genérico $\widehat{\pi_i}(1-\widehat{\pi_i})$:

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

$$V = \begin{bmatrix} \widehat{\pi_1}(1-\widehat{\pi_1}) & 0 & \dots & 0 \\ 0 & \widehat{\pi_2}(1-\widehat{\pi_2}) & \dots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & \dots & 0 & \widehat{\pi_n}(1-\widehat{\pi_n}) \end{bmatrix}$$

Uma vez ajustado o modelo, é necessário testar a sua significância. Uma das estratégias mais utilizadas é baseada na função de verosimilhança, particularmente no Teste da Razão de Verosimilhança.

No caso do modelo de regressão logística é através da Deviance e da Estatística de Qui-Quadrado de Pearson. A Deviance é análoga à soma de quadrados do modelo linear e é uma medida dos desvios no ajuste de um modelo de regressão logística aos dados em causa. É calculada através da

comparação dos valores observados e dos valores esperados, isto é, do modelo nulo (modelo em análise) com o modelo atual (modelo no qual o número de variáveis é igual ao número de observações).

O objetivo é testar simultaneamente se os coeficientes de regressão associados a $\boldsymbol{\beta}$ são todos nulos com exceção de $\boldsymbol{\beta_0}$.

A Deviance é definida por:

$$D = -2ln\left(\frac{\text{Função de máx. verosimilhança do modelo nulo}}{\text{Função de máx. verosimilhança do modelo atual}}\right) =$$

$$= -2\sum_{i=1}^{n}\left[y_i ln\left(\frac{\widehat{\pi_i}}{y_i}\right) + (1-y_i) ln\left(\frac{1-\widehat{\pi_i}}{1-y_i}\right)\right]$$

A hipótese a testar é então:

$$\mathrm{H_0}: \beta_1 = \cdots = \beta_0 = 0 \ \text{ vs } \ \mathrm{H_1}: \exists_{\mathrm{j=1,\dots,p}}: \beta_\mathrm{j} \neq 0$$

A estatística de teste é definida por:

$$G = D(\text{modelo sem as } p \text{ variáveis}) - D(\text{modelo com as } p\ variáveis)$$

$$\mathrm{G} = -2\ln\left(\frac{\text{modelo nulo}}{\text{modelo atual}}\right) \cap_{\text{sob } \mathrm{H_0}} \mathcal{X}^2_\mathrm{p}$$

(Hosmer & Lemeshow, 2000)

Se a hipótese nula for rejeitada conclui-se que pelo menos um dos coeficientes é estatisticamente diferente de zero. Antes da conclusão final dever-se-á então testar se cada um dos coeficientes é significativamente diferente de zero, sendo para isso realizado o Teste de Wald.

Como mencionado, o teste de Wald tem como objetivo averiguar se cada coeficiente é significativamente diferente de zero, isto é, se uma determinada variável independente apresenta uma relação estatisticamente significativa com a variável dependente:

$$H_0: \beta_j = 0 \ \ vs \ \ H_1: \beta_j \neq 0, \quad j = 0, \dots, p$$

Onde a estatística de teste é dada por:

$$W_j = \frac{\widehat{\beta_J}}{\widehat{SE}(\widehat{\beta_J})} \cap_{sob\ H_0} \mathcal{X}^2_{(1)}$$

Assim, a hipótese nula é rejeitada a um nível de significância α, se o valor observado da estatística de Wald for superior ao quantil de probabilidade $1-\alpha$.

Interpretação dos coeficientes estimados

Logo após o ajustamento do modelo e a avaliação à significância dos coeficientes estimados, procede-se à interpretação dos seus valores.

Intervalos de Confiança

Podem ser discutidos intervalos de confiança para os coeficientes estimados, para o logit e para as probabilidades logísticas calculadas.

Coeficientes

Os extremos de um intervalo de confiança de $100(1-\alpha)\%$ para os coeficientes estimados são obtidos através da seguinte maneira:

$$\left(\hat{\beta} \pm Z_{1-\frac{\alpha}{2}} \widehat{SE}(\hat{\beta})\right),$$

em que $Z_{1-\frac{\alpha}{2}}$: quantil de probabilidade $\left(1-\frac{\alpha}{2}\right)$ da distribuição Normal de valor médio nulo e variância unitária.

Logit

Para o logit de um modelo que contém p variáveis, o seu estimador será:

$$\hat{g}(x) = \widehat{\beta_0} + \widehat{\beta_1} x_1 + \widehat{\beta_2} x_2 + \cdots + \widehat{\beta_p} x_p$$

Uma alternativa que expressa o estimador do logit em linguagem matricial é:

$$\hat{g}(x) = x'\hat{\beta}$$

em que $\hat{\beta}' = \left(\widehat{\beta_0}, \widehat{\beta_1}, \widehat{\beta_2}, \dots, \widehat{\beta_p}\right)$ denota o estimador dos *p+1* coeficientes e o vetor $x' = \left(x_0, x_1, x_2, \dots, x_p\right)$ representa o valor da constante e o conjunto dos valores das p-covariáveis do modelo, onde $x_0 = 1$.

O estimador da variância do estimador do logit é dada por:

$$\widehat{VAR}[\hat{g}(x)] = \sum_{j=0}^{p} {x_j}^2 \widehat{VAR}\left(\widehat{\beta_j}\right) + \sum_{j=0}^{p} \sum_{k=j+1}^{p} 2 x_j x_k \widehat{COV}\left(\widehat{\beta_j}, \widehat{\beta_k}\right)$$

A partir da expressão do estimador da variância dos coeficientes estimados, $\widehat{VAR}\left(\hat{\beta}\right) = (X'VX)^{-1}$, vem que:

$$\widehat{VAR}[\hat{g}(x)] = x'\widehat{VAR}\left(\hat{\beta}\right)x = x'(X'VX)^{-1}x$$

Os valores a serem calculados tornam-se difíceis de calcular e por isso são usados softwares adequados que facilitam esses mesmos cálculos.

Assim, o intervalo de $100(1-\alpha)\%$ de confiança para o logit estimado é dado por:

$$\left(logit \ \pm Z_{1-\frac{\alpha}{2}} \ \widehat{SE}(\hat{\pi})\right)$$

em que $Z_{1-\frac{\alpha}{2}}$: quantil de probabilidade $\left(1-\frac{\alpha}{2}\right)$ da distribuição Normal de valor médio nulo e variância unitária.

Probabilidades Logísticas

A partir do intervalo de confiança do logit facilmente se consegue chegar ao intervalo de confiança da probabilidade logística estimada, $\hat{\pi}$, ou seja:

$$\left(\frac{e^{logit_{inf}}}{1+e^{logit_{inf}}}, \frac{e^{logit_{sup}}}{1+e^{logit_{sup}}}\right)$$

Por outro lado, o intervalo de confiança pode ser calculado da seguinte maneira:

$$P\left\{-q_{1-\frac{\alpha}{2}} < \frac{x^{*\prime}\hat{\beta} - x^{*\prime}\beta}{\sqrt{x^{*\prime}(X'VX)^{-1}x^{*}}} < q_{1-\frac{\alpha}{2}}\right\} = 1-\alpha$$

em ordem a $x^{*\prime}\beta$ tal que $x^{*\prime} = [1 \quad X_{i1} \quad \ldots \quad X_{ip}]$ em que i é a observação que queremos colocar em estudo e p é o número de variáveis do modelo.

(Hosmer & Lemeshow, 2000)

Seleção de Variáveis

O objetivo de qualquer um destes métodos é selecionar as variáveis que resultem no melhor modelo possível no contexto científico do problema.

De maneira a que este objetivo seja cumprido é necessário um plano básico para a seleção das variáveis para o modelo e um conjunto de métodos para avaliar a adequabilidade deste, tanto em termos das suas variáveis como do seu ajuste global.

Os critérios da inclusão de uma variável num modelo podem variar consoante o problema em causa. A abordagem tradicional para a construção de um modelo estatístico envolve a procura de um modelo mais parcimonioso na explicação dos dados.

A minimização do número de variáveis num modelo tem como objetivo uma maior estabilidade deste já que, quanto maior for o número de variáveis, maiores serão os erros padrões estimados e mais dependente se torna o modelo dos dados observados.

Assim, o processo de seleção de variáveis começa por uma análise univariada, ou seja, é feita uma análise às correlações entre cada uma das variáveis de maneira a que o modelo não tenha variáveis correlacionadas entre si. Em seguida, é feito um estudo de cada variável com a variável dependente de maneira a que o modelo seja constituído pelas variáveis independentes mais correlacionadas com a variável dependente. Depois desta análise estar completa, selecionar-se-ão as variáveis para a análise multivariada. A seleção é feita primeiramente através de técnicas como Árvores de Decisão, Qui-quadrado e R-quadrado. Posteriormente, o grau de importância de uma variável deve ser medido pelo p-value de Wald. Quanto menor for este valor mais importante será a variável para o modelo. Qualquer variável cujo p-value seja menor ou igual a 0.1 (ponto de entrada – p.e) deverá ser considerada candidata ao modelo, caso seja superior a 0.15 (ponto de saída – p.s) é candidata a sair.

Árvores de Decisão

Através da técnica de Árvores de Decisão obtém-se uma representação de uma tabela de decisão sob a forma de uma árvore. É uma abordagem comportamental que usa diagramas para sugerir alternativas e resultados de decisões assim como as probabilidades de ocorrência.

Uma árvore empírica representa uma segmentação dos dados. As regras são aplicadas sequencialmente, resultando numa hierarquia de segmentos à qual se chama árvore e cada segmento é um nó. O segmento original contém todo o conjunto de dados e é chamado de nó raiz da árvore. Um nó com todos os seus sucessores forma um ramo a partir do nó que o criou. Os nós finais chamam-se folhas. Para cada folha a decisão é tomada e aplicada para todas as observações que pertencem a essa folha. Neste caso a decisão é o valor previsto. (Breiman, 2001)

R-Quadrado

A seleção de variáveis através do R^2 é baseada numa Regressão Progressiva (Forward Stepwise) de Mínimos Quadrados que maximiza o valor R^2 do modelo. Relembre-se que o Método dos Mínimos Quadrados é uma técnica de otimização matemática que procura encontrar o melhor ajuste para um conjunto de dados, tentando minimizar a soma dos quadrados das diferenças entre o valor estimado e os dados observados, isto é, minimizar a soma dos quadrados dos resíduos. Este critério efetua uma rápida avaliação da variável e facilita o desenvolvimento de modelos preditivos com grandes volumes de dados (Universidade de Berkeley, 2011).

O coeficiente de correlação, R^2, de cada variável é calculado e comparado com o R^2 mínimo por defeito (0.005), embora este possa ser alterado conforme as necessidades de tratamento dos dados. Se o R^2 de uma variável for menor que o critério de corte então a variável é rejeitada, caso contrário é selecionada para ficar no modelo. Aumentar o ponto de corte pode originar a exclusão de mais variáveis enquanto que diminuí-lo pode causar o efeito contrário, ou seja, podem ser aceites mais variáveis no modelo. O R^2 varia entre 0 e 1. Quando toma o valor 0 conclui-se que a variável não tem relação linear com a variável explicativa. Caso seja igual a 1 significa que a variável explica toda a variabilidade da variável dependente (SAS, s.d.).

Qui-Quadrado

Esta técnica apenas é usada quando a variável explicativa é binária. Este teste mede a certeza dos valores observados poderem ser aceites no modelo.

A variável χ^2 é definida por

$$\chi_k^2 = \sum_{j=1}^{n} \frac{\left(O_j - E_j\right)^2}{E_j}$$

onde k representa os graus de liberdade (definido como a diferença entre o número de medidas realizadas e o número de restrições feitas aos valores das medidas), O_j é total de indivíduos observados na categoria j e E_j é o total de indivíduos esperados na categoria j.

Se $\chi^2 \leq n$ quer dizer que há uma boa relação entre as distribuições da variável independente com a variável resposta (Taylor, 1997).

Método para Seleção de Variáveis

Para selecionar o subconjunto de variáveis significativas de entre todas aquelas que estão disponíveis, utilizar-se-á um método de seleção Stepwise.

O método de seleção de variáveis Stepwise opera de forma iterativa partindo de um modelo sem nenhuma variável selecionada, e em cada iteração todos os termos (variáveis independentes), que ainda não estão incluídos no modelo e que constituem o conjunto de variáveis candidatas, são analisados. As variáveis cuja adição ao modelo implica um maior ganho no poder de precisão do modelo são adicionadas. Para além desta seleção, o algoritmo, em cada iteração, faz uma análise dos termos já selecionados com o intuito de detetar se algum destes pode ser excluído sem implicar uma grande perda de precisão no final.

Atualmente, praticamente todos os softwares têm uma opção para o Stepwise na regressão logística. A utilização deste método pode proporcionar uma maneira mais rápida e eficaz de selecionar um grande número de variáveis e de ajustar um certo número de equações da regressão logística simultaneamente.

Qualquer procedimento do Stepwise para a seleção ou supressão de variáveis de um modelo é baseado num algoritmo estatístico que verifica a importância das variáveis e as inclui ou exclui com base numa regra de decisão fixa. A importância de uma variável é definida pela significância estatística do seu coeficiente.

Considerem-se p variáveis independentes, todas tidas como importantes para explicar a variável resposta. Segue-se a descrição do método, com inclusão progressiva seguido de eliminação regressiva, com base na regra de decisão critica.

Passo 0. Em primeiro lugar, é ajustado um modelo contendo apenas os termos independentes e calcula-se o valor do logaritmo da verosimilhança, que se designa por L_0.

De seguida ajustam-se vários modelos univariados, um para cada uma das p variáveis independentes e calcula-se o valor do logaritmo da verosimilhança para cada um desses modelos. Designa-se por $L_j^{(0)}$ o logaritmo da verosimilhança para o modelo que contém a variável x_j no passo 0. O subscrito j refere-se à variável incluída no modelo e o sobrescrito refere-se ao passo. Esta notação será utilizada ao longo do processo do Stepwise para controlar tanto o número de passos como o número de variáveis no modelo.

Para o modelo que contém a variável x_j versus o modelo que contém apenas os termos constantes, o valor do teste da razão das verosimilhanças é dado por:

$$G_j^{(0)} = 2\left(L_j^{(0)} - L_0\right)$$

Neste caso, o p-value é determinado por:

$$P\left(\mathcal{X}_v^2 > G_j^{(0)}\right) = p_j^{(0)}, \text{ onde } v = \begin{cases} k-1\,, x_j \text{ policotómica} \\ 1, \text{variável contínua} \end{cases}$$

A variável considerada estatisticamente mais importante é aquela a que corresponde a um p-value menor.

Seja $p_{e_1}^{(0)} = min_j\left(p_j^{(0)}\right)$, em que p_e é o p-value correspondente à variável mais importante. Se $p_{e_1}^{(0)} < p_e$ passa-se ao passo seguinte, caso contrário acaba aqui. O subscrito e_1 é utilizado para indicar que a variável é candidata a entrar no passo 1.

Passo 1. Ajustamento do modelo que contém apenas a variável que corresponde ao menor p-value.

Seja x_{e_1} essa variável e $L_{e_1}^{(1)}$ o logaritmo da verosimilhança para este modelo.

Dado que o modelo já contém a variável x_{e_1} , das $p-1$ variáveis, é necessário determinar quais as mais importantes. Ajustam-se $p-1$ variáveis modelos contendo a variável x_{e_1} e $x_j, j = 1, \dots, p\ e\ j \neq e_1$. Designa-se $L_{e_1 j}^{(1)}$ pelo logaritmo da verosimilhança que contém x_{e_1}e x_j.

Seja então a estatística do qui-quadrado do modelo dada por:

$$G_j^{(1)} = 2\left(L_{e_1 j}^{(1)} - L_{e_1}^{(1)}\right)$$

Tal como no passo anterior, $p_j^{(1)}$ é o p-value correspondente à estatística apresentada.

Seja x_{e_2} a variável que deu origem ao menor p-value neste passo. Então vem que $p_{e_2}^{(1)} = min_j\left(p_j^{(1)}\right)$. Assim, se $p_{e_2}^{(1)} < p_e$ prossegue-se para o passo seguinte, caso contrário acaba neste passo.

Passo 2. Como tem sido feito, inicia-se com o ajustamento do modelo contendo as variáveis x_{e_1} e x_{e_2}.

Uma vez que se trata, simultaneamente, de uma seleção de variáveis progressiva e regressiva, é possível que ao incrementar a variável x_{e_2}, a variável x_{e_1} possa ter deixado de ser importante. Então está presente a selecção regressiva.

Seja $L_{-e_j}^{(2)}$ o valor do logaritmo da verosimilhança do modelo retirando a variável x_{e_2}, e a estatística do qui-quadrado do modelo sem a variável x_{e_2} e do modelo com ambas as variáveis, definida por:

$$G_{-e_j}^{(2)} = 2\left(L_{e_1e_2}^2 - L_{-e_j}^{(2)}\right)$$

Tal que $p_{-e_j}^{(2)}$ é o p-value correspondente.

Seja x_{s_2} a variável que foi excluída do modelo cujo p-value que lhe corresponde é definido como $p_{s_2}^{(2)} = max\left(p_{-e_1}^{(2)}, p_{-e_2}^{(2)}\right)$. Para decidir se a variável x_{s_2} será ou não removida do modelo, compara-se o valor de p com p_s, sendo p_s o p-value para a saída de variáveis do modelo.

Assim, se $p_s^{(2)} < p_s$, a variável x_{s_2} será removida do modelo. Caso contrário mantem-se.

Passando à seleção progressiva são ajustados $p-2$ modelos contendo x_{e_1}, x_{e_2} e $x_j,\ j = 1,2,\dots,p,\ j \neq e_1, e_2$. Calcula-se o logaritmo da verosimilhança para cada um dos modelos encontrados para cada um dos $p-2$ modelos tal como na iteração 1, determina-se a estatística de teste da razão de verosimilhança para estes novos modelos vs o modelo contendo apenas as variáveis x_{e_1} e x_{e_2}, com o cálculo dos respetivos p-value.

Sendo x_{e_3} a variável que corresponde ao menor p-value encontrado, prossegue-se para o passo seguinte se $p_{e_3}^{(2)} < p_e$, caso contrário esta é a última iteração.

⋮

Passo n. Neste passo todas as p variáveis já se encontram no modelo, isto é, todas as variáveis constituintes do modelo têm p-value inferiores ao valor de p_s e todas as variáveis que não foram incluídas no modelo têm p-value superiores a p_e.

Assim o modelo que resulta deste passo conterá todas as variáveis que foram consideradas estatisticamente importantes de acordo com os valores de p_e e p_s.

(Hosmer & Lemeshow, 2000)

Análise do Ajustamento do Modelo

A avaliação de um modelo é essencial para medir até que ponto este é capaz de fazer corretamente as previsões. Essa avaliação é feita separando o conjunto de dados disponíveis no conjunto de treino e no conjunto de validação. O modelo é construído com base no conjunto de treino e depois aplicado no conjunto de validação para depois se compararem os valores de validação com a previsão. Assim o modelo é avaliado e as incertezas de previsão são estimadas.

Existem várias medidas de avaliação dos modelos e que interpretam a sua performance na classificação dos dados, tais como a Matriz de Confusão e a Curva de Lift descritas seguidamente.

Matriz de confusão

A matriz de confusão é uma tabela de visualização dos resultados em que cada linha representa as instâncias previstas de uma classe enquanto cada coluna representa as instâncias reais de uma classe. Torna-se uma ferramenta útil para analisar a qualidade do classificador no reconhecimento de exemplos de diferentes categorias. (Han & Kamber, 2006)

Em geral costuma-se chamar valores positivos aos valores (observados ou previstos) iguais à unidade e valores negativos àqueles que são nulos. Os valores positivos que são previstos como tal chamam-se verdadeiros positivos (VP) e os que são previstos erradamente como negativos são denominados como falsos negativos (FN). Analogamente, os verdadeiros valores negativos podem ser previstos corretamente como sendo negativos, isto é, os verdadeiros negativos (VN), ou podem ser previstos como positivos e, nesse caso, denominam-se falsos positivos (FP) (Alpuim, 2016). Depois dos pares de valores observados/previstos estarem todos classificados os resultados são apresentados no que se chama de Matriz de Confusão, que apresenta os totais de pares em cada uma das categorias tal como ilustra a tabela:

	Previstos		
Observados	Sim	Não	Total
Sim	VP	FN	VP + FN
Não	FP	VN	FP+ VN
Total	VP + FP	FN + VN	nº de observações

Tabela 1: Matriz de Confusão

Uma medida que marca, no contexto da classificação em Data Mining, o limiar a partir do qual a previsão se altera tem o nome de Threshold. Cada previsão tem uma confiança associada e são selecionados como positivos os casos cuja confiança é superior ao valor de Threshold. A variação deste valor provoca o reajuste das previsões e nova classificação dos registos em teste. Esta medida varia entre 0 e 1, e assumir algum destes valores extremos classifica todos os registos como negativos ou positivos respetivamente. A diminuição do valor de Threshold promove a classificação dos registos como positivos e diminui a classificação dos registos como negativos, o que implica que aumentem os registos corretamente classificados como positivos, ou seja, os Verdadeiros Positivos, mas também os registos erradamente classificados como positivos, ou seja, os Falsos Positivos. Ao contrário, diminuem os registos corretamente classificados como negativos, isto é, os Verdadeiros Negativos e diminuem os registos incorretamente classificados como negativos, isto é, os Falsos Negativos. Idealmente era

desejável ter um elevado número de registos corretamente identificados e um baixo número de registos classificados incorretamente. Não sendo possível, a variação do Threshold permite avaliar qual o equilíbrio em termos de classificação de registos.

Em suma, a diminuição do valor de Threshold faz aumentar o número dos Verdadeiros Positivos e reduzir o número dos Falsos Negativos ainda acompanhado com um aumento de Falsos Positivos.

Com o objetivo de apreciar a qualidade do modelo, isto é, a sua percentagem de acerto, definem-se as seguintes probabilidades:

1. **Sensibilidade (S):** probabilidade de uma observação ser classificada como positiva dado que é, de facto, positiva.

$$S = \frac{VP}{VP + FN}$$

2. **Especificidade (E):** probabilidade de uma observação ser classificada como negativa dado que ela é, de facto, negativa.

$$E = \frac{VN}{VN + FP}$$

A percentagem de indivíduos corretamente classificados é dada por $\frac{VP+VN}{FN+VN} \times 100$.

A partir da matriz de confusão podemos também quantificar os erros de tipo I (falsos positivos) e de tipo II (falsos negativos) que se baseiam nas perdas incorridas em cada caso e assim, segundo o tamanho amostral, permitem definir o nível de significância ideal que minimiza a combinação linear dos erros de decisão.

Erro Tipo I: é a probabilidade de rejeitar H_o sabendo que H_o é verdadeira, isto é, quando no modelo os Clientes são considerados vinculados quando na realidade eles dizem não o ser.

$$Erro\ Tipo\ I = \frac{FP}{VN + FP}$$

Erro Tipo II: é a probabilidade de não rejeitar H_o sabendo que H_o é falsa, isto é, quando no modelo os Clientes não são considerados como vinculados e na realidade consideram o B o seu Banco Principal. Este é o erro considerado mais grave.

$$Erro\ Tipo\ II = \frac{FN}{FN + VP}$$

Definimos assim o **Erro Ponderado** e o **Erro Total**:

$$Erro\ Ponderado = (Erro\ Tipo\ I \times 0.5) + (Erro\ Tipo\ II \times 0.5)$$

$Erro\ Total = \frac{FP+FN}{n}$, onde *n* é o número total da amostra

Finalmente, quantificamos ainda o nível de **Acuidade ou Precisão**, ou seja, a probabilidade dos dados previstos (total de acertos), estarem próximos dos verdadeiros valores da amostra, ou seja, é o grau de proximidade entre os valores previstos e os valores reais.

$$Precisão = \frac{VN + VP}{n}$$

Curva de Lift

A curva de Lift permite avaliar o desempenho de um modelo de regressão binária. É uma das métricas mais importantes na avaliação de modelos preditivos. Neste caso vai ser estudado o Lift Acumulado que indica quantas vezes, num determinado nível de rejeição, o modelo é melhor do que a seleção aleatória (Modelo Aleatório) (Coppock, 2002). Mais precisamente, é o rácio entre a proporção de Clientes vinculados e a resposta da variável dependente quando analisada toda a população.

Flow-chart

Para facilitar a compreensão do modelo construído apresenta-se um esquema que resume os passos que vão ser realizados para a construção do Modelo de Vinculação de um Cliente a um Banco:

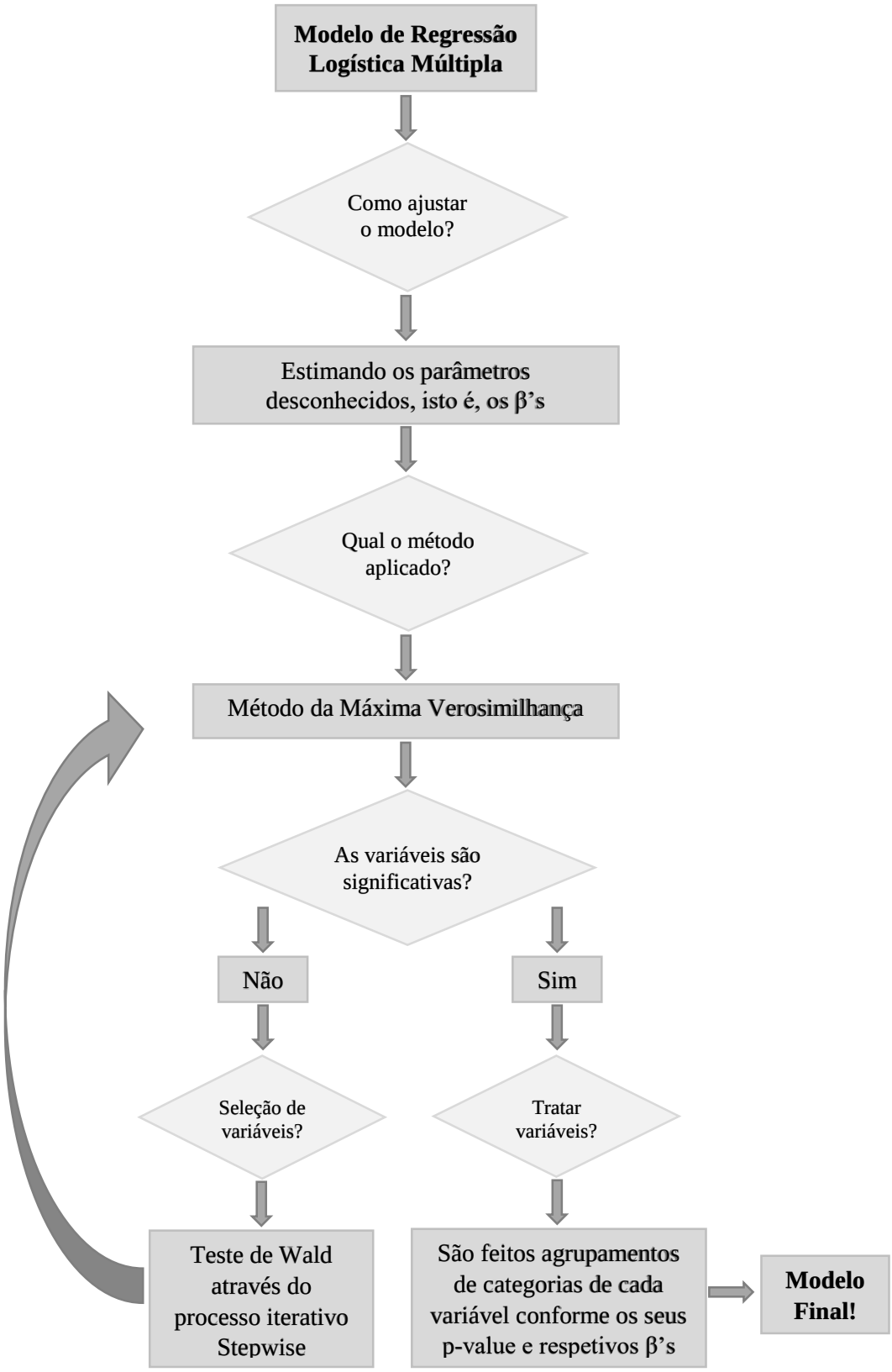

Figura 1: Flow Chart do Modelo de Vinculação de um Cliente a um Banco

Capítulo II

Neste capítulo determina-se a vinculação de um Cliente a um Banco aplicando as metodologias teóricas desenvolvidas anteriormente. Desta forma é possível determinar quais as características, que em conjunto, levam a concluir que um Cliente é ou não é vinculado.

Objetivo

Tendo como objetivo o estudo de um modelo estatístico diferenciador da vinculação pretende-se, em função do perfil de cada Cliente, analisar a forma como esta se expressa na sua relação com o Banco, isto é, caracterizar a posse e utilização dos principais produtos financeiros de cada Cliente que levam a concluir acerca da vinculação do Cliente ao Banco.

O software utilizado na construção do modelo de regressão logística e na sua análise foi o SAS Enterprise Miner Client 13.2.

Comparativamente a outras técnicas, por exemplo à Regressão Linear, a Regressão Logística distingue-se pela variável resposta ser categórica, fator que permite comparar ainda com outros métodos como Árvores de Decisão, Redes Neuronais, entre outros. Uma vez que a finalidade é ter todo o conjunto de características que levam o Cliente a ser Vinculado ao Banco e ainda ter a capacidade interpretativa do impacto marginal de cada variável na vinculação do Cliente, a Regressão Logística é a melhor metodologia a ser aplicada neste estudo.

Para a recolha de informação foi elaborado um questionário para ser aplicado a alguns Clientes do Banco B, em que a pergunta é "Considera o Banco B o seu Banco Principal?". A resposta é sim caso o Cliente tenha exclusivamente o Banco B como Banco Principal ou então mais do que um banco e mesmo assim o considere como principal, e caso contrário é não, constituindo assim a variável resposta.

Para a seleção do Universo de Modelação foram inquiridos Clientes num estudo regular de satisfação, nas suas edições de 2011 a 2016, os quais têm uma relação regular com o Banco. No total, a amostra é constituída por cerca de 6000 Clientes.

O resultado desta predição estatística sobre a vinculação do Cliente ao Banco tem o propósito de melhorar a informação utilizada em outros estudos de satisfação e em indicadores de previsão relativamente à quota da relação bancária de cada Cliente com o Banco, entre outros.

Variáveis em estudo

As variáveis podem ser classificadas em dois tipos: variáveis qualitativas (categóricas) e variáveis quantitativas (numéricas).

As variáveis qualitativas são definidas por várias categorias, ou seja, representam uma classificação dos indivíduos. Podem ser nominais ou ordinais, mas não é possível aplicar operações aritméticas sobre elas. Já as variáveis quantitativas podem ser ordenadas e submetidas a operações aritméticas. Existem variáveis quantitativas discretas (assumem apenas valores enumeráveis e somente fazem sentido valores inteiros) e variáveis quantitativas contínuas (assumem qualquer valor real, ou seja, até valores fracionais fazem sentido).

Quanto às variáveis temos que, para além da Target (constituída pelas respostas aos questionários feitos a cada Cliente) que é binária, existem, no total, 52 variáveis (binárias, intervalares, nominais e ordinais).

A variável resposta (target) é binária e, portanto, um resultado igual a 1 indica que o Cliente considera o Banco B como Banco Principal, enquanto que um valor zero indica que o Banco B não é considerado o seu Banco Principal. Aproximadamente 75% da amostra são Clientes vinculados ao Banco e 25% não são vinculados.

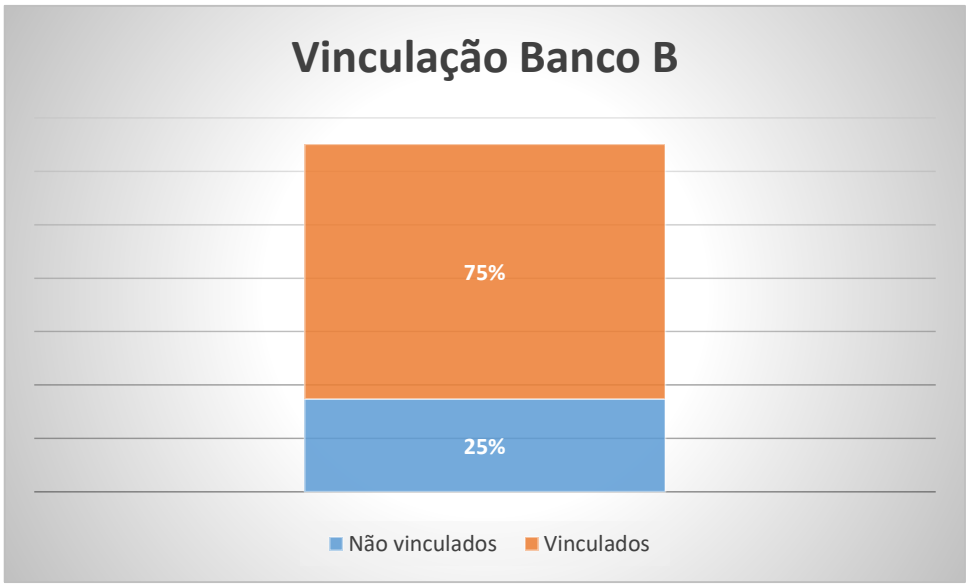

Gráfico 1: Distribuição da Vinculação dos Clientes do Banco B

As variáveis do modelo distribuem-se da seguinte forma:

1. Dados Sociodemográficos
2. Dados Socioeconómicos
3. Níveis de utilização do Banco B via cartões por tipologia transacional
4. Dados da Conta e Posse de Produtos e Serviços
   - **4.1** Posse de Produtos no Banco B
   - **4.2** Posse de Produtos de Crédito na Banca
5. Tipologia de Investimento

Em termos de formulação, $X_{i,j}$ é a variável do modelo em que i representa o tema (de acordo com os 5 temas mencionados acima) a que a variável pertence e j é o número da variável no tema i.

Capítulo III

Tratamento de dados do modelo

Cada vez mais os modelos preditivos são utilizados pelo mercado a fim de ajudarem as empresas na mitigação de riscos, expansão de carteiras, prevenção de fraudes, retenção de Clientes, entre outros.

Esta fase baseia-se na preparação dos dados e no tratamento de variáveis. Ou seja, após a amostragem dos dados, o passo seguinte é explorá-los e ver os seus comportamentos e possíveis agrupamentos. A exploração ajuda a refinar o modelo, pois a maioria dos dados com que estamos a trabalhar podem não ser ideais para proceder ao desenvolvimento do modelo. Alguns dos problemas que surgem são, por exemplo, *missing values* ou valores em falta, valores não numéricos, dados errados ou inconsistentes.

Alguns dos métodos mais utilizados no tratamento de *missing values* envolvem a sua remoção ou substituição. O procedimento mais comum é tratar os dados eliminando os *missings* e construir o modelo apenas com os dados completos, mas os resultados podem-se tornar enviesados se a amostra não for completamente representada. Em contrapartida, há quem não inclua no modelo as variáveis com *missings*, mas isto pode levar a que o poder preditivo do modelo seja inferior ao que seria obtido se todas as variáveis fossem testadas. Poder-se-á estar a prescindir de variáveis que na realidade são importantes para modelar o fenómeno e sendo este o caso, o modelo provavelmente nunca produzirá resultados tão precisos quanto estariam ao seu alcance, caso as referidas variáveis de *input* fossem utilizadas. Com vista a resolver este problema, foram desenvolvidos métodos que têm como objetivo preencher os *missings*, o que origina uma análise de dados mais completa com todos os indivíduos e variáveis.

Uma segunda opção para o problema consiste em manualmente examinar os exemplos com valores omissos e introduzir um valor provável ou que se considere razoável tendo em conta o perfil do registo. Este método pode resultar caso o número de registos com *missings values* não seja muito grande. No entanto, há que ter em atenção que caso não haja um valor óbvio e plausível para cada registo, pode-se correr o risco de estar a introduzir informação errónea no conjunto de dados.

Outra opção para a resolução do problema traduz-se no preenchimento automático dos campos com uma boa estimativa do seu valor. Existem diversas formas de produzir esta estimativa, sendo que a mais óbvia e simples consiste em adotar uma medida de tendência central, como a média, a moda ou a mediana.

Outra forma, mais sofisticada, mas também mais trabalhosa, consiste em desenvolver um modelo preditivo que, com base nos registos completos e nas variáveis disponíveis, nos dê uma estimativa para os valores omissos.

Finalmente, existem situações onde o facto de não haver valores em determinados campos pode ser um indicador importante. Por isso, interessa desenvolver uma codificação específica para que o modelo possa captar esta característica do conjunto de dados. Existe, no entanto, um perigo associado a esta abordagem e que se traduz na possibilidade de o modelo associar erroneamente o valor omisso com o *output*. Existem formas de tentar controlar estes perigos, quando utilizamos ferramentas que não produzem um modelo interpretável do seu funcionamento como a análise de sensibilidade. No entanto, não existe substituto para o raciocínio lógico, e não havendo explicação para determinados resultados proporcionados por um modelo, deve-se ter o máximo cuidado na sua aplicação.

À medida que a dimensionalidade do espaço de *input* aumenta, também aumenta a probabilidade de ocorrência de correlações espúrias, sem qualquer aderência à realidade e que acontecem por acaso. Facilmente se compreende que ao utilizar 10 variáveis a probabilidade de por acaso haver uma

correlação sem significado real é muito menor do que se o espaço de *input* tiver 200 variáveis. (Bação, s.d.)

Para além disto, existem ainda valores que "caem" fora da região normal de interesse do espaço de *input*. Esses valores são definidos como *outliers*. Podem representar situações fora do vulgar que, no entanto, estão corretas. Mas podem também corresponder a medições incorretas e que por isso se tornam enganadoras, podendo prejudicar a performance do modelo. Genericamente, pode-se dizer que os *outliers* são valores que aparecem nas caudas do histograma, sendo assim relativamente fáceis de detetar. Outra forma de os identificar consiste no recurso à teoria estatística, ou mais especificamente à distribuição normal procedendo-se ao cálculo da média e desvio-padrão para cada variável de input, definindo como *outliers* todos os valores que se encontrarem fora do intervalo do desvio padrão a partir da média.

Na regressão logística, quando há muitas variáveis independentes que podem ser potencialmente incluídas no modelo, é uma boa prática proceder à análise bivariada entre estas e a variável dependente. A Análise Bivariada permite a análise simultânea de duas (ou mais) variáveis. Neste caso permite estabelecer relações entre um *input* e a target, ou seja, determinar se as diferenças entre a distribuição de duas variáveis são estatisticamente significativas com o objetivo de encontrar influências, causalidades ou coincidências.

A Discretização dos dados é muito utilizada em conjunto com técnicas de exploração de dados para comprovar a influência de certas variáveis no resultado obtido. Pode ser definida como o processo de transformação de uma variável contínua numa variável discreta.

Como as variáveis intervalares são medições contínuas de uma escala linear pode surgir a preocupação sobre os efeitos não lineares destas variáveis com a variável dependente. A não linearidade pode ser uma simples curva (quadrática ou exponencial) ou algo mais complicado onde o relacionamento não é monótono e, portanto, é feita uma discretização da variável onde este tipo de variáveis passa a ordinais. As variáveis ordinais são muito úteis para avaliar qualidades que não podem ser quantificadas de forma objetiva, como acontece no caso, por exemplo, da situação profissional. Concretamente neste caso, as novas variáveis ordinais foram obtidas a partir da discretização de variáveis intervalares pela partição da amplitude por um número finito de classes.

Deve ainda ser levado em conta que muitos problemas que normalmente ocorrem com variáveis contínuas como aprendizagem mais lenta, ausência de generalização e geração de muitos relacionamentos com pouco poder preditivo, também ocorrem se a quantidade de possíveis valores ou a quantidade de intervalos de uma variável discreta for muito grande. Por isso a discretização deve resultar preferencialmente em variáveis discretas com poucos intervalos.

Assim a discretização em si pode também ser encarada como uma forma de descoberta de conhecimento, onde podem ser revelados os valores críticos de um domínio contínuo. Os intervalos de discretização não devem esconder padrões de relacionamento entre as variáveis, mas devem ser escolhidos com cuidado ou podem ser perdidos potenciais resultados.

Uma vez que o padrão de relacionamento das variáveis independentes com a variável dependente poderá não ser linear e é distinto de variável para variável, foi estipulado o processo de discretização de todas as variáveis, com o objetivo de facilitar a interpretação do perfil do Cliente através das suas características, ou seja, simplificar a interpretação do impacto que a alteração de uma característica pode causar no nível de vinculação de um Cliente. Caso contrário o estudo a ser feito teria de ser variável a variável linearizando as relações entre a variável dependente com a variável independente em causa. O tratamento de *missings* e de *outliers* acaba por ser um benefício da discretização. Para o estudo das variáveis não se tornar tão "artificial" com estas suposições, o método de discretização aplicado sugere que os *missings* constituam uma categoria de maneira a que contribuam com informação e assim se consigam tirar conclusões sobre o perfil de cada Cliente.

Tratamento da Multicolinearidade e Seleção de Variáveis

Devido à quantidade elevada de informação que o modelo possui, é necessário, através de alguns critérios, reduzir o tamanho da amostra, mas assegurando a representatividade das características da população em estudo. Para tal, é feito um tratamento de multicolinearidade e consequente seleção de variáveis através de técnicas como o qui-quadrado, R^2, árvores de decisão, coeficiente de Gini, entropia e uma regressão logística grosseira.

Exemplos

Para uma melhor perceção do que é realmente feito na discretização seguem-se alguns exemplos:

Exemplo 1 – Variável 24 dos Dados da Conta e Posse de Produtos e Serviços - $X_{4,24}$

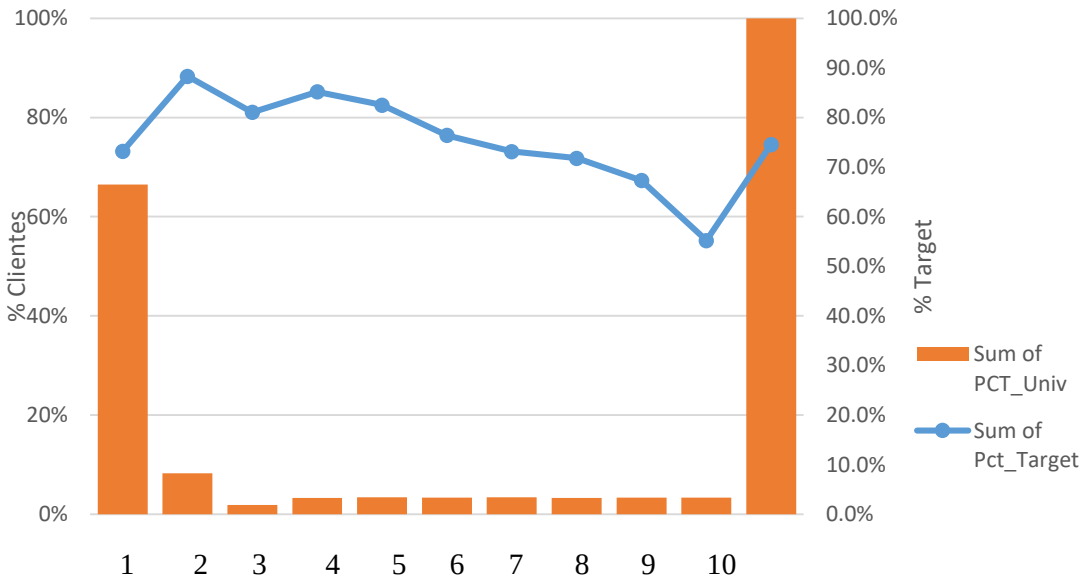

Gráfico 2: Comportamento da Variável 24 dos Dados da Conta e Posse de Produtos no processo de Discretização

1. A primeira categoria da variável é constituída por *missings*, isto é, por informação que o Banco B não consegue dispôr relativamente aos seus Clientes.
2. Com todas as observações da variável, excetuando os *missings*, vão ser construídos até 10 intervalos.
3. A variável passa a ser distribuída pelos intervalos formados, ou seja, neste caso, do segundo ao décimo intervalo representados no gráfico estão todas as observações que constituem agora 100% da variável.
4. No processo de contagem da frequência do primeiro intervalo, o total de observações é superior a 10% e, como a discretização não faz a partição dos elementos com a mesma característica, a massa deste intervalo ultrapassa os 10%. Este intervalo conta quase como 2 intervalos e meio de massa igual a 10%, portanto vai ser procurada uma observação que

constitua um novo intervalo que somado com o anterior formem 3 intervalos. Assim surge o segundo intervalo que é o terceiro representado no gráfico. A partir daqui são construídos mais 7 intervalos com igual frequência de valores.

5. Voltando a agregar os *missings*, que tinham sido isolados, a variável sofre um novo ajuste. Por isso é que os valores finais são os que estão representados no gráfico 2.

Exemplo 2 – Variável 1 dos Dados da Conta e Posse de Produtos e Serviços - $X_{4,1}$

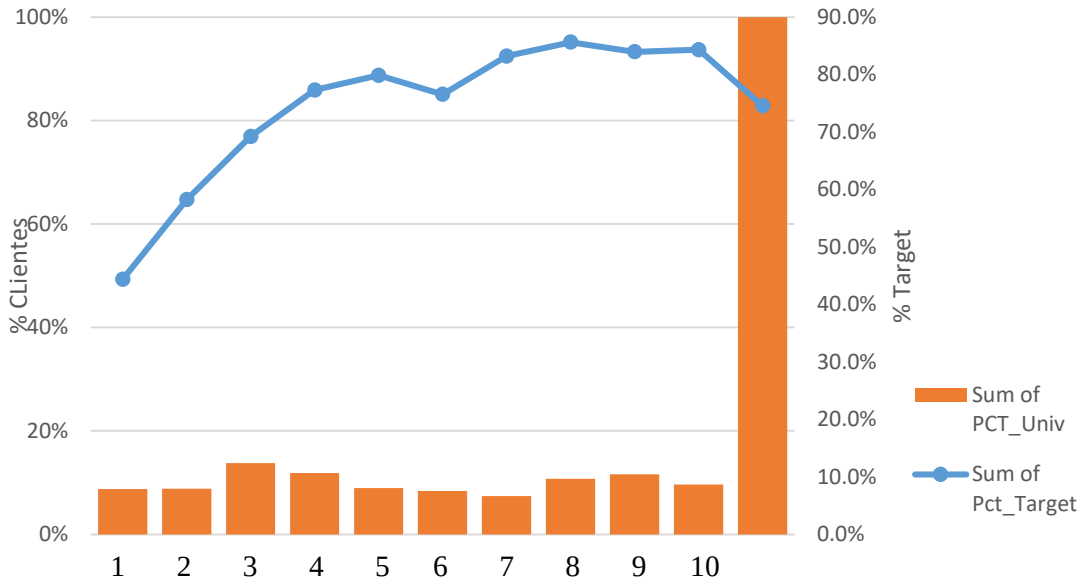

Gráfico 3: Comportamento da Variável 1 dos Dados da Conta e Posse de Produtos e Serviços no processo de Discretização

1. Esta variável não tem *missings* por isso é feita a categorização da variável imediatamente.
2. Primeiramente são construídos os intervalos 3, 4, 8 e 9 por terem uma frequência de observações superior a 10%.
3. Depois de contados os intervalos com maior massa são construídos os 6 restantes intervalos de maneira a terem igual frequência de valores tal como está representado no gráfico 3.

Em suma, a fase de exploração e tratamento de dados abordou a discretização das variáveis acompanhada de uma análise bivariada.

No final deste processo, depois das variáveis intervalares passarem a ordinais, os *missings* existentes nas variáveis intervalares acabaram por ser tratados como benefício da discretização e encarados como uma própria categoria, contribuindo desta forma com informação para o desenvolvimento do modelo.

Nesta altura as variáveis apresentam algumas categorias que podem sofrer algumas alterações, isto é, podem ser agrupadas. Por exemplo, um Cliente que tenha uma responsabilidade de longo prazo com o Banco B e que só esteja em vias de concluir a sua amortização será equiparado a quem não tem essa responsabilidade e por isso faz sentido, do ponto de vista de vinculação, agrupar estes dois tipos de Clientes. Estes agrupamentos são fundamentados na técnica do R^2 que será explicado mais adiante.

Caso as variáveis binárias tivessem valores em falta, estes iriam ser tratados logo de seguida, mas como não apresentam, o output do "tratamento de *missings*" acaba por não apresentar qualquer tipo de resultado.

Depois dos dados estarem todos tratados o modelo está pronto para passar à fase seguinte.

Multicolinearidade do modelo

Em Regressões é muito comum existirem problemas como a Multicolinearidade, ou seja, as variáveis independentes possuem relações lineares exatas ou aproximadamente exatas entre elas (Schaefer, 1986). Se não existem relações lineares entre as covariáveis do modelo, isto é, se forem ortogonais, podem ser realizadas inferências sobre os efeitos das covariáveis na variável dependente. Infelizmente, na maioria dos casos, as covariáveis não são ortogonais e conclui-se que o problema da multicolinearidade existe. Consequentemente, a inferência baseada nessas estimativas pode estar seriamente comprometida. Quando o R^2 entre duas variáveis é bastante elevado (acima de 70%), mas nenhum dos coeficientes da regressão é estatisticamente significativo, pode-se dizer que existe um indício de multicolinearidade.

Alguns procedimentos para o tratamento da multicolinearidade são aumentar o tamanho da amostra, usar informações *a priori* sobre o valor da estimativa dos parâmetros ou ainda excluir as variáveis colineares. (Cotidiano, 2013)

A multicolinearidade é tratada através da análise de 3 casos:

1. Ausência de Multicolinearidade – ocorre quando a correlação entre as variáveis explicativas é nula (situação ideal);
2. Multicolinearidade Perfeita – a correlação entre as variáveis explicativas é igual a 1 ou -1. O cálculo das estimativas dos parâmetros é matematicamente impossível nestas circunstâncias;
3. Multicolinearidade Imperfeita – a correlação entre as variáveis pertence aos intervalos]0,1[e]-1, 0[que é o caso mais comum e o qual exige um tratamento da multicolinearidade.

As variáveis que têm uma grande correlação entre elas transmitem essencialmente a mesma informação por isso, neste caso, vão ser escolhidas apenas algumas variáveis, acabando por excluir outras, o que origina uma seleção de variáveis. Essa seleção é feita também através das correlações entre cada variável com a target, ou seja, quanto maior a correlação entre cada uma das covariáveis com a variável independente melhor, pois significa que contêm a mesma informação o que é bom para o modelo que está a ser construído.

Em primeiro lugar procedeu-se à extração da informação de cada variável. Algumas correlações eram de facto elevadas e, portanto, de maneira a não perder qualquer informação potencialmente importante, foram criados dois grupos de variáveis (grupo de posse e grupo de valor) para serem submetidos a uma posterior seleção. Ambos os grupos possuem as mesmas variáveis com exceção das que são correlacionadas, ou seja, um tem maioritariamente variáveis que indicam o que o Cliente tem no Banco B e o outro quanto é que o Cliente tem nesse mesmo Banco.

Depois de formados os grupos segue-se a seleção de variáveis aplicando diferentes técnicas. As que surgirem mais vezes significa que são as que seguem para a modelação.

Seleção de Variáveis do modelo

Árvores de Decisão

Algumas variáveis dos nós da árvore de decisão são selecionadas em relação à sua importância relativa e as mais importantes para o modelo são as que registam valores de significância elevados. Para este modelo, o método de divisão e escolha das melhores variáveis intervalares teve como critério a função Fisher em que o p-value do teste F está associado à variância de cada nó. Em relação às variáveis

ordinais, o método de seleção da divisão dos nós é baseado na entropia e, neste caso, foram aplicados 3 métodos nas árvores de decisão que especificam as regras de divisão e a escolha das melhores variáveis quanto às variáveis nominais: Qui-Quadrado, Entropia e Índice de Gini. (SAS, s.d.)

Qui-Quadrado

O teste de Qui-Quadrado de Pearson é a base deste algoritmo. Constitui um método estatístico extremamente eficiente para a segmentação, ou crescimento de uma árvore. Encontra-se explicado mais à frente, uma vez que também é uma técnica usada numa outra prática de seleção de variáveis.

Uma das vantagens deste algoritmo é o facto de parar o crescimento da árvore antes de ocorrer o problema de *overfitting* (quando o modelo estatístico se ajusta demasiado ao conjunto de dados).

Como as árvores de decisão são um pouco extensas vão ser interpretados apenas os primeiros nós.

Por exemplo, no grupo de variáveis que representam a posse dos Clientes, o nó de nível superior da árvore de decisão mostra que, dos 6341 Clientes que responderam ao questionário, 75% dizem que consideram o Banco B como Banco Principal e 25% não consideram. Sob este nó, a variável que corresponde à Quantidade de Levantamentos é a que melhor caracteriza o modelo, isto é, tem mais significância registada nesta técnica. Para uma quantidade de levantamentos superior a 3 ou na categoria de valores omissos é mais provável que os Clientes se considerem vinculados ao Banco B do que se a quantidade de levantamentos for inferior a 3. É uma diferença de 84.35% de Clientes supostamente vinculados para 57.03%. Uma grande variedade de técnicas de divisão foi desenvolvida ao longo do tempo para avaliar se esta diferença é estatisticamente significativa e se os resultados são precisos e reprodutíveis. Na Figura 2, a diferença entre uma quantidade de levantamentos superior ou inferior a 3 é significativa.

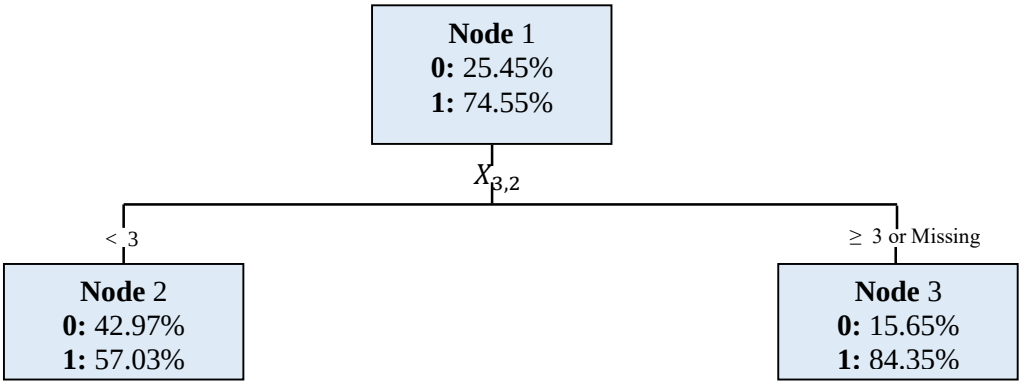

Figura 2: 3 primeiros nós da Árvore de Decisão do grupo das variáveis que representam a Posse dos Clientes do Banco B selecionadas

A importância de cada uma das variáveis selecionadas por esta técnica, para este grupo de variáveis, está indicada na seguinte tabela:

Variáveis	Importância
$X_{3,2}$	1
$X_{1,1}$	0,5516
$X_{1,2}$	0,4936
$X_{4,2}$	0,3983
$X_{4,3}$	0,3913
$X_{2,1}$	0,2276
$X_{4,6}$	0,2244
$X_{1,4}$	0,2129
$X_{4,13}$	0,1527
$X_{4,7}$	0,1261

Tabela 2: Importância das variáveis que representam as Posses dos Clientes do Banco B selecionadas segundo o Qui-Quadrado nas Árvores de Decisão

No grupo de variáveis que retrata o valor que os Clientes possuem, a variável mais significativa já não é a mesma do que no grupo anterior. Como pode ser verificado na Figura 3, o primeiro corte foi feito na covariável Número de Movimentos de Débito. Assim, é mais provável que em mais do que 3 movimentos de débito ou em valores omissos os Clientes sejam mais vinculados ao Banco B (84.87%) do que em menos do que 3 movimentos de débito (58.92%).

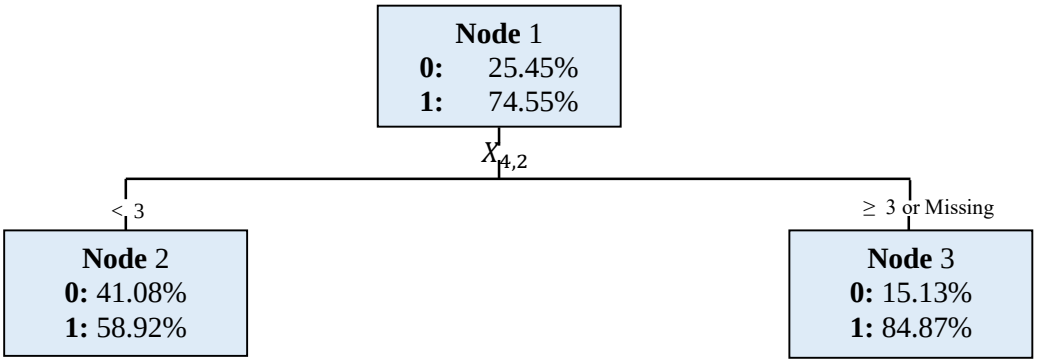

Figura 3:3 primeiros nós da Árvores de Decisão do grupo das variáveis que representam o Valor que os Clientes possuem no Banco B selecionadas

Para este grupo de variáveis, a importância de cada uma delas, está classificada da seguinte forma:

Variáveis	Importância
$X_{4,2}$	1
$X_{1,2}$	0,6481
$X_{1,1}$	0,3739
$X_{2,2}$	0,3016
$X_{4,24}$	0,2782
$X_{4,3}$	0,2673
$X_{1,4}$	0,2212
$X_{4,6}$	0,218
$X_{4,30}$	0,1998
$X_{5,5}$	0,1768
$X_{4,28}$	0,1757
$X_{4,29}$	0,1476
$X_{4,1}$	0,1464
$X_{5,6}$	0,1229

Tabela 3: Importância das variáveis que representam o Valor que os Clientes possuem no Banco B selecionadas segundo o Qui-Quadrado nas Árvores de Decisão

Entropia

A entropia é uma medida aplicável à partição de um espaço de probabilidade, ou seja, mede quanto esse espaço é homogéneo. Suponha-se um conjunto de vários exemplos S, e um conjunto de n classes $C = \{C_1, C_2, \dots, C_n\}$, sendo p_i a probabilidade da classe C_i em S. Então a entropia do conjunto S é a homogeneidade deste e é definida por

$$Entropia = -\sum_{i=1}^{c} p_i \log_2 p_i$$

Quanto maior a entropia maior a desordem. Quanto menor for a entropia, menor a profundidade das árvores geradas e menor o número de nós e ramificações.

As variáveis selecionadas no grupo de posse neste processo foram:

Variáveis	Importância
$X_{3,2}$	1
$X_{1,1}$	0,5418
$X_{1,2}$	0,424
$X_{4,2}$	0,4171
$X_{1,3}$	0,3712
$X_{4,3}$	0,332
$X_{4,6}$	0,2321
$X_{2,1}$	0,2276
$X_{4,17}$	0,1987

$X_{2,2}$	0,1375
$X_{4,4}$	0,0956
$X_{4,14}$	0,0722

Tabela 4: Importância das variáveis do grupo de Posse selecionadas segundo a Entropia nas Árvores de Decisão

No grupo de valor as variáveis selecionadas foram:

Variáveis	Importância
$X_{4,2}$	1
$X_{1,2}$	0,6503
$X_{1,1}$	0,5012
$X_{1,3}$	0,4047
$X_{2,2}$	0,3852
$X_{4,24}$	0,2792
$X_{4,3}$	0,2606
$X_{4,6}$	0,2582
$X_{3,3}$	0,1392
$X_{2,1}$	0,1141

Tabela 5: Importância das variáveis do grupo de Valor selecionadas segundo a Entropia nas Árvores de Decisão

Índice de Gini

Segundo o critério de Gini, o grau de impureza num dado nó é dado por

$$G(N) = 1 - \sum_{I=1}^{L} p^2(I|N)$$

Este critério é definido por uma variável nominal com L categorias, onde $p(I|N)$ é a probabilidade *a priori* da classe I se formar no nó N. Cada variável pode ser usada diversas vezes ao longo do processo de crescimento da árvore. Deste modo, este índice contabiliza a proporção de observações em cada classe da variável dependente num nó relativamente ao total (nó raiz). O índice de Gini assume o seu valor mínimo quando num nó correspondente a uma partição da variável dependente, apenas existem observações pertencentes a uma classe. (Rodrigues, 2005)

Após ser aplicado o critério de Gini as variáveis selecionadas no grupo de posse foram:

Variáveis	Importância
$X_{3,2}$	1
$X_{1,1}$	0,5776
$X_{4,2}$	0,4643
$X_{1,2}$	0,4597
$X_{4,3}$	0,3668
$X_{1,3}$	0,3255
$X_{2,1}$	0,2865
$X_{4,6}$	0,1588

$X_{2,2}$	0,1375
$X_{4,1}$	0,1363
$X_{4,4}$	0,0926
$X_{4,14}$	0,0722

Tabela 6: Importância das variáveis do grupo de Posse selecionadas segundo o Índice de Gini nas Árvores de Decisão

No grupo de valor as variáveis selecionadas foram:

Variáveis	Importância
$X_{4,2}$	1
$X_{1,2}$	0,6497
$X_{1,1}$	0,5074
$X_{1,3}$	0,4533
$X_{2,2}$	0,308
$X_{4,3}$	0,2856
$X_{4,24}$	0,2789
$X_{3,3}$	0,2519
$X_{4,6}$	0,2185
$X_{2,1}$	0,2128
$X_{4,30}$	0,2003
$X_{5,5}$	0,1822
$X_{4,28}$	0,1661
$X_{4,15}$	0,1245
$X_{4,7}$	0,1043

Tabela 7: Importância das variáveis do grupo de Valor selecionadas segundo o Índice de Gini nas Árvores de Decisão

R-Quadrado

As variáveis selecionadas para o modelo por esta técnica, no grupo de posse, foram as seguintes, como podemos ver na tabela 8:

Variáveis	gl	R^2	F	p-value	SQ_{TOT}	EQM
$X_{4,2}$	3	0,1069	252,74	<,0001	128,574156	0,169577
$X_{1,1}$	5	0,0248	36,12	<,0001	29,801574	0,165004
$X_{3,2}$	4	0,0155	28,75	<,0001	18,650335	0,162161
$X_{1,4}$	2	0,0108	40,59	<,0001	13,000771	0,160157
$X_{4,6}$	1	0,0103	78,19	<,0001	12,371637	0,158227
$X_{4,3}$	1	0,0064	49,32	<,0001	7,744944	0,157027
$X_{1,2}$	2	0,0074	28,44	<,0001	8,85343	0,155676
$X_{4,1}$	4	0,0056	10,82	<,0001	6,696029	0,154715
$X_{4,14}$	1	0,0046	36,14	<,0001	5,55975	0,153859
$X_{4,17}$	1	0,0031	24,54	<,0001	3,762367	0,153288
$X_{4,8}$	1	0,0024	18,55	<,0001	2,83549	0,152863
$X_{1,3}$	1	0,0014	10,76	0,001	1,642659	0,152627
$X_{2,2}$	4	0,0012	2,36	0,0512	1,439015	0,152496
$X_{4,22}$	1	0,0010	7,79	0,0053	1,186297	0,152332

Tabela 8: Variáveis do gripo de Posse selecionadas segundo o R-Quadrado nas Árvores de Decisão

em que a Soma de Quadrados Total (SQ_{TOT}) representa uma medida de variação ou desvio da média, isto é, a dispersão de cada indivíduo relativamente à média total. Esse valor é calculado como a soma dos quadrados das diferenças das médias, e inclui a soma dos quadrados dos fatores e aleatoriedade ou erro (resíduos):

$$\sum (y - \bar{y})^2 = \sum (\hat{y} - \bar{y})^2 + \sum (y - \hat{y})^2$$

$$SQ_{TOT} = SQ_{Trat} + SQ_E$$

O Erro Quadrático Médio (EQM) é definido como sendo a média da diferença entre o valor do estimador e do parâmetro ao quadrado, isto é, é usado para indicar o quão distante, em média, o conjunto de estimativas está do único parâmetro a ser estimado:

$$EQM = \frac{\sum (y - \hat{y})^2}{N - k} = \frac{SQ_E}{N - k}$$

tal que N é o número de observações e k o número de graus de liberdade (gl) (Gomes, 2015).

As variáveis selecionadas no grupo de valor foram as seguintes:

Variáveis	gl	R^2	F	p-value	SQ_{TOT}	EQM
$X_{4,2}$	3	0,1069	252,74	<,0001	128,574156	0,169577
$X_{1,1}$	5	0,0248	36,12	<,0001	29,801574	0,165004
$X_{4,30}$	3	0,0162	40,20	<,0001	19,538778	0,161995
$X_{1,4}$	2	0,0125	47,27	<,0001	15,095875	0,15966
$X_{5,5}$	4	0,0114	21,79	<,0001	13,732489	0,15759
$X_{4,6}$	1	0,0086	66,54	<,0001	10,37824	0,155973
$X_{1,2}$	2	0,0086	33,60	<,0001	10,373739	0,154381
$X_{4,3}$	1	0,0075	58,97	<,0001	9,021657	0,152978
$X_{4,24}$	5	0,0065	10,30	<,0001	7,822405	0,15186
$X_{4,27}$	4	0,0043	8,63	<,0001	5,214882	0,15113
$X_{4,1}$	4	0,0032	6,46	<,0001	3,893777	0,150608
$X_{5,2}$	7	0,0024	2,70	0,0086	2,838116	0,150325
$X_{3,3}$	2	0,0018	7,26	0,0007	2,179429	0,150027
$X_{4,29}$	4	0,0015	3,02	0,0167	1,811891	0,149834
$X_{2,2}$	4	0,0013	2,65	0,0315	1,586744	0,149677
$X_{4,23}$	2	0,0013	5,07	0,0063	1,517109	0,149483
$X_{4,22}$	1	0,0008	6,81	0,0091	1,016697	0,149345
$X_{1,3}$	1	0,0010	7,93	0,0049	1,182703	0,149181
$X_{4,28}$	2	0,0007	2,63	0,0721	0,784351	0,149104
$X_{4,8}$	1	0,0005	4,07	0,0437	0,606268	0,149031

Tabela 9: Variáveis do grupo de Valor selecionadas segundo o R-Quadrado nas Árvores de Decisão

Qui-Quadrado

O teste do Qui-Quadrado foi desenvolvido por Karl Pearson e completado por Ronald Fisher (Stigler, 2011).

Este teste é usado para analisar o grau da associação entre as variáveis independentes e a variável resposta, ou seja, se o modelo encontrado explica bem os dados.

A estatística de teste é dada por:

$$\chi^2 = \sum_{i=1}^{n} r(y_i, \hat{\pi}_i)^2 \cap \chi^2_{n-p-1}$$

onde os resíduos de Pearson são definidos por $r_i = r(y_i \hat{\pi}_i) = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}}$, $i = 1,2, \dots, n$.

As variáveis selecionadas para o modelo no grupo de posse foram as seguintes:

Variáveis	Importância Relativa
$X_{3,2}$	1.0
$X_{1,1}$	0.5083
$X_{4,2}$	0.4872
$X_{1,2}$	0.4514
$X_{2,1}$	0.4101
$X_{4,3}$	0.3320
$X_{1,3}$	0.2875
$X_{4,6}$	0.2725
$X_{4,1}$	0.2016
$X_{4,17}$	0.1323
$X_{4,22}$	0.0843
$X_{2,2}$	0.0424

Tabela 10: Importância das varáveis do grupo de Posse selecionadas segundo o Qui-Quadrado

As variáveis selecionadas no grupo de valor foram as seguintes:

Variáveis	Importância Relativa
$X_{4,2}$	1.0
$X_{1,2}$	0.6557
$X_{1,1}$	0.4499
$X_{1,3}$	0.3586
$X_{2,2}$	0.3198
$X_{2,1}$	0.3052
$X_{4,3}$	0.2882
$X_{4,24}$	0.2815
$X_{4,6}$	0.2206
$X_{4,30}$	0.2021
$X_{4,26}$	0.1979

$X_{5,2}$	0.1859
$X_{3,3}$	0.1615
$X_{5,5}$	0.1511
$X_{4,8}$	0.1051
$X_{4,23}$	0.0989
$X_{4,28}$	0.0884
$X_{4,1}$	0.0860

Tabela 11: Importância das variáveis do grupo de Valor selecionadas segundo o Qui-Quadrado

Regressão Logística

Tendo em conta toda a amostra é feita uma primeira análise de uma regressão logística de maneira a ver quais as variáveis a serem selecionadas através do método Stepwise.

As variáveis selecionadas no grupo de posse correspondem ao seguinte resultado:

Variável	gl	Estatística de Wald (X^2)	Pr > ChiSq
$X_{4,8}$	1	6,45	0,0111
$X_{4,16}$	1	5,3891	0,0203
$X_{4,3}$	1	26,766	<,0001
$X_{4,9}$	1	22,1533	<,0001
$X_{4,21}$	1	5,2577	0,0219
$X_{2,1}$	8	30,8228	0,0002
$X_{4,4}$	1	7,0215	0,0081
$X_{4,6}$	1	26,7826	<,0001
$X_{4,17}$	1	27,9757	<,0001
$X_{4,13}$	1	4,2789	0,0386
$X_{4,15}$	1	8,6036	0,0034
$X_{4,22}$	1	13,546	0,0002
$X_{1,4}$	2	91,5053	<,0001
$X_{1,1}$	8	106,546	<,0001
$X_{1,2}$	3	64,8257	<,0001
$X_{4,1}$	4	18,823	0,0009
$X_{4,2}$	4	24,0519	<,0001
$X_{3,2}$	5	112,9647	<,0001

Tabela 12: Variáveis do grupo de Posse selecionadas segundo o Stepwise na Regressão Logística

As variáveis selecionadas no grupo de valor correspondem ao seguinte resultado:

Variável	gl	Estatística de Wald (X^2)	Pr > ChiSq
$X_{4,8}$	1	4,1348	0,042
$X_{4,3}$	1	36,6367	<,0001
$X_{4,21}$	1	5,0128	0,0252
$X_{2,1}$	8	25,3376	0,0014
$X_{4,6}$	1	24,4884	<,0001
$X_{4,22}$	1	14,7676	0,0001
$X_{4,23}$	4	14,8748	0,005
$X_{4,24}$	4	37,6451	<,0001
$X_{4,27}$	4	15,3326	0,0041
$X_{4,28}$	3	11,9659	0,0075
$X_{4,29}$	5	11,9285	0,0358
$X_{4,30}$	3	15,545	0,0014
$X_{1,4}$	2	98,2333	<,0001
$X_{1,1}$	8	99,0796	<,0001
$X_{1,2}$	3	68,6266	<,0001
$X_{4,1}$	4	17,3837	0,0016
$X_{4,2}$	4	94,8179	<,0001
$X_{3,3}$	2	19,1196	<,0001
$X_{5,5}$	5	24,3219	0,0002

Tabela 13: Variáveis do grupo de Valor selecionadas segundo o Stepwise na Regressão Logística

Depois de aplicadas todas as técnicas de seleção de variáveis são verificadas quais as variáveis independentes que são selecionadas mais do que uma vez no total de todas as técnicas. Naturalmente que existem algumas variáveis selecionadas que são correlacionadas entre si e portanto, mais uma vez, são analisadas as correlações entre as variáveis e, principalmente, a correlação de cada uma das variáveis com a variável dependente de maneira a que as variáveis selecionadas sejam aquelas mais correlacionadas com a variável explicativa.

Foram excluídas, através da seleção dos métodos descritos, as variáveis $X_{4,12}$, $X_{4,18}$, e $X_{4,10}$. Pelas correlações entre as variáveis, saem as que são menos correlacionadas com a target, o que acontece com grande parte das variáveis de posse que dão lugar às variáveis de valor. Como exemplo, a variável de $X_{4,17}$ foi eliminada e ficou selecionada a variável de valor $X_{4,29}$. É excluída a variável $X_{4,25}$ e fica no modelo a variável $X_{4,26}$ por ser mais informativa do ponto de vista de negócio. Uma variável muito correlacionada com outras variáveis independentes, inclusive com a variável $X_{3,2}$, era a variável $X_{3,3}$ que acabou por ser excluída do modelo.

Após ter sido reduzido o espaço de input o modelo segue para a fase de modelação onde se vão decidir quais as variáveis finais do estudo da Vinculação de um Cliente com o Banco B.

Modelação

Esta fase tem como objetivo determinar as variáveis finais para o modelo em estudo e estudar as significâncias das variáveis. O método de Regressão que vai analisar a amostra é o Stepwise. É um processo semi-automático de construção do modelo onde se acrescentam ou eliminam sucessivamente variáveis, ou seja, é um método que concilia os métodos de análise Backward e Forward. (Hosmer&Lemeshow, 2000)

A análise através do Backward começa num modelo com todas as covariáveis aceites e vai eliminando as que não são significativas (através dos testes de Wald), enquanto que através do Forward parte do zero e, passo a passo, testa a inclusão das variáveis e a qualidade do modelo resultante.

A metodologia Stepwise parte do modelo nulo (como no Forward) e verifica em cada inclusão de uma nova variável a importância das covariáveis já presentes no modelo sendo possível a remoção de alguma variável (tal como no Backward) caso esta se tenha tornado desnecessária.

O valor de significância utilizado é 10%, para que o modelo construído não seja composto por poucas variáveis, e a significância das variáveis é trabalhada através do agrupamento das suas categorias até todas as variáveis do modelo serem significativas.

Na tabela 14 encontram-se as características das variáveis resultantes no modelo final:

Variável	Categorias	$\widehat{\beta}$	$\widehat{S}_{\widehat{\beta}}$	Estatística de Wald (X^2)	p-value	$e^{\widehat{\beta}}$
Intercept		1,7994	0,1337	181,22	<,0001	6,046
$X_{4,3}$	0	-0,2703	0,0509	28,22	<,0001	0,763
	1	0,2703	0,0509	28,22	<,0001	1,31
$X_{4,6}$	0	-0,358	0,0492	52,97	<,0001	0,699
	1	0,358	0,0492	52,97	<,0001	1,431
$X_{4,23}$	0	-0,2732	0,0927	8,69	0,0032	0,761
	1	0,4838	0,1185	16,67	<,0001	1,622
	2	-0,2106	0,1315	2,56	0,1094	0,81
$X_{4,24}$	0	0,2933	0,0727	16,29	<,0001	1,341
	1	-0,2933	0,0727	16,29	<,0001	0,746
$X_{4,27}X_{4,28}$	1	-0,4027	0,1004	16,1	<,0001	0,669
	2	0,6966	0,153	20,74	<,0001	2,007
	3	-0,2939	0,1284	5,24	0,0221	0,745
$X_{4,29}X_{4,30}$	1	-0,1299	0,0816	2,53	0,1115	0,878
	2	0,7456	0,101	54,54	<,0001	2,108
	3	-0,6157	0,1017	36,62	<,0001	0,54
$X_{1,1}$	0	0,2829	0,0542	27,2	<,0001	1,327
	1	-0,2829	0,0542	27,2	<,0001	0,754
$X_{1,2}$	0	-0,6433	0,0742	75,08	<,0001	0,526
	1	-0,0623	0,0717	0,76	0,3845	0,94
	2	0,7056	0,1058	44,49	<,0001	2,025
$X_{4,1}$	0	-0,1842	0,0526	12,27	0,0005	0,832
	1	0,1842	0,0526	12,27	0,0005	1,202
$X_{4,2}$	0	-0,2904	0,0598	23,56	<,0001	0,748
	1	0,2904	0,0598	23,56	<,0001	1,337
$X_{2,1}$	0	-0,1671	0,0515	10,52	0,0012	0,846
	1	0,1671	0,0515	10,52	0,0012	1,182
$X_{3,2}$	0	-0,4378	0,0539	66,01	<,0001	0,645

	1	0,4378	0,0539	66,01	<,0001	1,549
$X_{5,5}$	0	-0,1322	0,0523	6,39	0,0115	0,876
	1	0,1322	0,0523	6,39	0,0115	1,141
$X_{1,3}$	0	1,1897	0,1512	61,91	<,0001	3,286
	1	-0,3133	0,0918	11,64	0,0006	0,731
	2	-0,5792	0,0845	47,01	<,0001	0,56
	3	-0,2972	0,0783	14,39	0,0001	0,743

Tabela 14: Variáveis finais do modelo segundo a Regressão Logística

De maneira a acertar as significâncias das variáveis, foram criadas duas variáveis em função de outras que já existiam. Foi o caso da variável $X_{4,27}X_{4,28}$ que é a junção das variáveis $X_{4,27}$ com $X_{4,28}$ e da variável $X_{4,29}X_{4,30}$ que é a junção das variáveis $X_{4,29}$ com $X_{4,30}$.

Há p-values muito elevados como é o caso da categoria 1 da variável $X_{1,2}$ mas como estes resultados provêm do teste do β ser ou não igual a zero, e como -0.0623 é praticamente zero a significância desta categoria foi aceite. Relativamente à 2ª categoria da variável $X_{4,23}$ e à 1ª categoria da variável $X_{4,29}X_{4,30}$, as suas significâncias são aproximadamente 10% e portanto são aceites no modelo.

Como se pode verificar na tabela seguinte, todas as variáveis finais são significativas:

Variáveis	Estatística de Wald (t)	p-value
$X_{4,3}$	28,2218	<,0001
$X_{4,6}$	52,9658	<,0001
$X_{4,23}$	20,7428	<,0001
$X_{4,24}$	16,2916	<,0001
$X_{4,27}X_{4,28}$	23,9533	<,0001
$X_{4,29}X_{4,30}$	56,6215	<,0001
$X_{1,1}$	27,1988	<,0001
$X_{1,2}$	75,2686	<,0001
$X_{4,1}$	12,2722	0,0005
$X_{4,2}$	23,563	<,0001
$X_{2,1}$	10,5168	0,0012
$X_{3,2}$	66,0108	<,0001
$X_{5,5}$	6,3851	0,0115
$X_{1,3}$	68,1458	<,0001

Tabela 15: Significância das variáveis finais do modelo

O R^2 do modelo final é igual a 0.2265, ou seja, 22.65%.

Avaliação do modelo

A avaliação de um modelo é uma das fases mais importantes do processo de tratamento de dados. Para fazer esta avaliação foram usadas medidas como o lift e as matrizes de confusão.

O que se pretende é que os números totais de falsos positivos (FP) e de falsos negativos (FN) sejam os menores possíveis. Estes valores podem ser alterados alterando o ponto de corte. Contudo, se, por exemplo, o ponto de corte for aumentado, diminui o número de FP mas aumenta o número de FN. Inversamente, se diminuirmos o valor do ponto de corte, diminui também o número de FN mas aumenta o número de FP. Ou seja, por modificação do ponto de corte, não é possível diminuir o número de falsos positivos sem aumentar os falsos negativos bem como inversamente.

Tendo em conta os Erros de Tipo I e de Tipo II, Especificidade, Sensibilidade e Acuidade ou Precisão vamos decidir qual o ponto de corte (Threshold) do modelo. De acordo com os dados da tabela 16, ficou estabelecido um corte de 65%, ou seja, com vista a não promover um Cliente não Vinculado quando este se declara como Vinculado e para não ser classificado como Vinculado quando na realidade não se considera como tal, a partir dos 65% é estabelecido o nível de Vinculação dos Clientes do Banco B.

Corte	TIPO I	Especificidade	TIPO II	Sensibilidade	Acuidade ou Precisão	ERRO TOTAL	ERRO PONDERADO
50,0%	56,5%	43,5%	7,6%	92,4%	80,0%	20,0%	32,0%
55,0%	52,2%	47,8%	10,1%	89,9%	79,2%	20,8%	31,1%
60,0%	46,7%	53,3%	12,6%	87,4%	78,7%	21,3%	29,6%
65,0%	41,9%	58,1%	15,9%	84,1%	77,5%	22,5%	28,9%
70,0%	35,1%	64,9%	21,2%	78,8%	75,3%	24,7%	28,1%
75,0%	29,4%	70,6%	27,5%	72,5%	72,0%	28,0%	28,4%
80,0%	21,8%	78,2%	35,9%	64,1%	67,7%	32,3%	28,9%
85,0%	14,5%	85,5%	46,8%	53,2%	61,4%	38,6%	30,7%
90,0%	7,9%	92,1%	63,0%	37,0%	51,0%	49,0%	35,5%

Tabela 16: Avaliação do modelo final

A partir dos 65% há um aumento significativo (mais de 5%) do Erro Tipo II e uma queda, também significativa (acima de 1%), na Acuidade ou Precisão e ainda na Sensibilidade, e portanto, 65% é considerado o corte do modelo.

Através do gráfico 4 seguinte consegue-se ver a evolução dos erros:

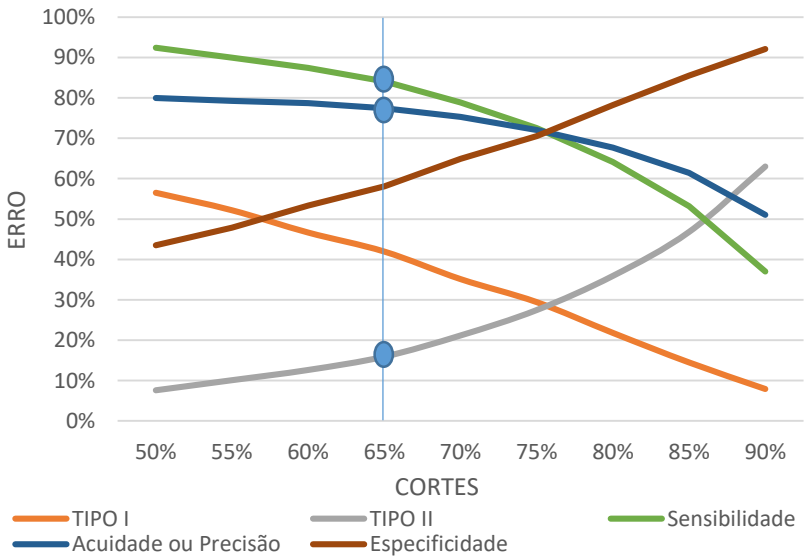

Gráfico 4: Evolução dos eros que avaliam o modelo final

A Matriz de Confusão confronta a classificação real dos Clientes com a prevista para o modelo. Para classificar um Cliente relativamente à sua vinculação é necessário definir a probabilidade de corte, que como vimos em cima, foi decidida para 65%.

Real	Previsão 0	1	Total
0	14,78%	10,68%	25,45%
	58,05%	41,95%	
	55,48%	14,55%	
1	11,86%	62,69%	74,55%
	15,91%	84,09%	
	44,52%	85,45%	
Total	26,64%	73,36%	100%

Legenda:
%total
%linha
%coluna

Gráfico 5: Matriz de Confusão do modelo final

Com uma propensão à vinculação igual ou superior a 0.65 (ponto de corte), indicar-se-ia que 73.36% da população é Vinculada ao Banco B. Capturar-se-iam cerca de 84.09% dos Clientes realmente Vinculados. Nessa ação, 85.45% de Clientes vinculados, apenas 74.55% dos Clientes são realmente vinculados, o que corresponde a um lift de 1.1462 (0.8545/0.7455=1.1462).

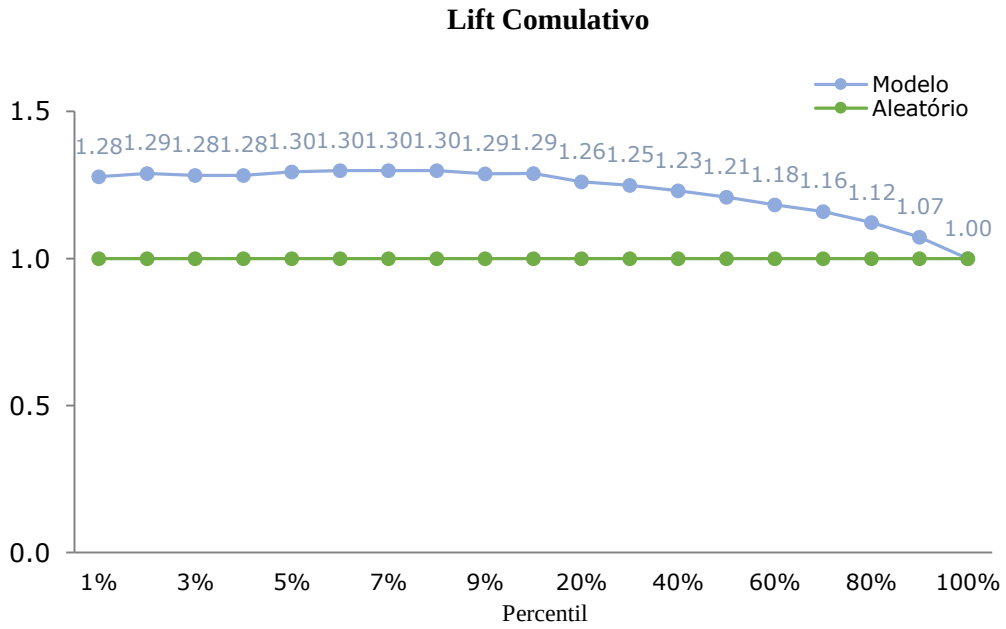

Gráfico 6: Lift Comulativo do modelo final

O lift é considerado a alavancagem do modelo. Apresenta-se como um indicador da eficiência do modelo face a um hipotético modelo aleatório, que mede quantas vezes mais é preferível a sua utilização, isto é, representa o ganho da utilização do modelo face à seleção aleatória de Clientes. Para 1% do universo, o modelo consegue capturar 1.28 vezes mais Clientes vinculados do que face ao acaso.

Análise do Perfil de Clientes

Depois de avaliar o modelo já se pode olhar para os Perfis dos Clientes do Banco B de maneira a tirar conclusões sobre o seu nível de Vinculação (como pode ser consultada na tabela que se encontra em anexo). Através de um grupo de variáveis sociodemográficas (SD), socioeconómicas (SE), comportamentais (COMP) e de carteira (CT) ir-se-á comparar o perfil dos Clientes identificados pelo modelo como Vinculados com o dos Clientes que declaram a sua vinculação.

A análise de perfil dos Clientes compreende o mapeamento das suas características e a identificação do nível de vinculação atual.

Validação do modelo

De maneira a validar o modelo que foi construído, usou-se a vaga de 2016, ou seja, dados do inquérito aos Clientes realizado em 2016.

Com o mesmo ponto de corte, isto é, cutoff = 65% obtemos a seguinte matriz de confusão (Tabela 17):

Teste (2016)

Real	Previsão 0	Previsão 1	Total
0	14,62%	10,34%	24,96%
	58,57%	41,43%	
	58,93%	13,75%	
1	10,19%	64,85%	75,04%
	13,58%	86,42%	
	41,07%	86,25%	
Total	24,81%	75,19%	

Modelação e Validação

Real	Previsão 0	Previsão 1	Total
0	14,78%	10,68%	25,45%
	58,05%	41,95%	
	55,48%	14,55%	
1	11,86%	62,69%	74,55%
	15,91%	84,09%	
	44,52%	85,45%	
Total	26,64%	73,36%	

Tabela 17: Matriz de Confusão de Teste ao modelo final e Matriz de Confusão de Modelação e Validação do modelo final

Tipo I	41,43%
Tipo II	13,58%
Total	20,53%
Precisão	79,47%
Corte =	0,65
Lift =	1,15

Tipo I	41,95%
Tipo II	15,91%
Total	22,54%
Precisão	77,46%
Corte =	0,65
Lift =	1,1462

Tabela 18: Erros, Precisão, Corte e Lift do Teste ao modelo final e da Modelação e Validação do modelo final

Através dos dados de 2016 foi testada e confirmada a vinculação dos Clientes que declararam que o Banco B é o seu Banco Principal. O modelo desenvolvido aplicado à vaga de 2016 apresenta resultados que não diferem dos previstos. Para 75.19% da população capturar-se-iam cerca de 86.42% dos Clientes realmente Vinculados. Incorporaria cerca de 86.25% dos Clientes Vinculados.

Podemos ainda concluir que a capacidade do modelo prever corretamente os Clientes não Vinculados, ou seja, a especificidade é de 58.57%; a capacidade do modelo prever corretamente os Clientes Vinculados, ou seja, a sensibilidade, é de 86.42%

A probabilidade dos dados previstos estarem próximos dos reais é dado pela precisão que é de 79.47% (14.62% + 64.85%).

Capítulo IV

Neste capítulo fazer-se-á uma pequena discussão acerca do modelo que foi desenvolvido mencionando algumas limitações e até mesmo alternativas a metodologias que foram praticadas.

Termo Independente

Se for calculada a probabilidade da Vinculação de um Cliente considerando todas as variáveis independentes iguais a zero, ou seja, quando calculado:

$$\frac{e^{\beta_0}}{1+e^{\beta_0}} = \frac{e^{1.7994}}{1+e^{1.7994}} = 0.858076 \simeq 86\%$$

conclui-se que um Cliente que não tenha quaisquer tipos de produtos financeiros e que ocupa as categorias mais baixas das restantes variáveis é assumido automaticamente como Vinculado ao Banco B.

Discretização

Como vimos a discretização atua como um método de transformação de valores contínuos em discretos. Neste caso, variáveis intervalares passam a ser variáveis ordinais. Para o processo de discretização ser efetuado, matematicamente, existirá uma "rutura epistemológica" pois a transformação de um raciocínio intuitivo, de natureza contínua, num raciocínio computacional, de natureza discreta, é um obstáculo epistemológico (Setti, 2009).

Este processo é essencial para o pré-processamento dos dados, não só porque alguns métodos não trabalham valores contínuos, mas também porque melhora o ponto de vista interpretativo e conclusivo (Liu, Hussain, Tan & Dash, 2002) e diminui o espaço de dados tornando o processo computacional mais rápido (Mittal & Cheong, 2002).

Existem muitas maneiras de se realizar o processo de discretização. Uma dessas maneiras consiste numa aprendizagem supervisionada e a outra é através de uma aprendizagem não supervisionada. A aprendizagem supervisionada consiste na construção de classes tendo em conta a variável explicativa. O sistema tem de determinar a descrição para cada classe, ou seja, o conjunto de propriedades que são comuns à variável dependente. Depois da descrição de cada classe estar concluída é possível formular uma regra de classificação que pode ser utilizada para prever a classe de um objeto que não tenha sido considerado aquando da aprendizagem. Pelo contrário, uma aprendizagem não supervisionada é efetuada com base em observações e descobertas, isto é, não são definidas classes, pelo que o sistema de Data Mining necessita de observar os exemplos e reconhecer os padrões por si próprio de onde resultam um conjunto de descrições de classes na base de dados. (Tipos de Data Mining, 2017)

Neste caso, aborda-se uma discretização que é uma junção de "Equal Width Discretization" com "Equal Frequency Discretization" e em que se tem em conta o facto de não serem formados diferentes intervalos com os mesmos dados, ou seja, trata-se de uma aprendizagem supervisionada. Portanto, a qualidade supervisionada ou não supervisionada de um método de discretização é um critério importante e que deve ser levado em consideração (Muhlenbach & Rakotomalala, 2005).

Como tal, a escolha de um método de discretização particular depende da sua complexidade algorítmica (algoritmos complexos terão mais tempo de computação e não serão adequados a conjuntos de dados muito grandes), da sua eficiência e da sua combinação com o método de aprendizagem.

A discretização pode causar relações não lineares entre as variáveis, por exemplo, jovens e séniores podem apresentar níveis de Vinculação idênticos. Segundo Frank & Witten (1999), as variáveis idade e, adaptando a este caso, a Vinculação, não apresentam uma relação de linearidade e é por esta razão que muitas vezes se opta por fazer a discretização de variáveis mesmo que o método computacional consiga trabalhar com variáveis contínuas. A interpretação dos resultados é muito mais fácil quando a variável está disposta por categorias e para tal as relações entre os atributos precisam de ser muito bem definidas, caso contrário os resultados podem ser mal interpretados (Camilo & Silva, 2009). No entanto esta abordagem não é muito viável. É preciso ter em atenção que se um indivíduo pertencer ao limite inferior de uma categoria vai ter um comportamento completamente diferente de um indivíduo que esteja no limite superior da categoria seguinte. Consequentemente gerar-se-á um maior numero de parâmetros a estimar (β's) e, por sua vez, muitos mais erros associados.

Segundo Wang et al. (2008), neste tratamento de dados existem problemas estatísticos, de precisão dos dados e padronizações, técnicos e organizacionais. A manipulação dos dados e a análise das informações de maneira tradicional torna-se inviável devido ao grande volume de dados.

Assim, a categorização de uma variável pode pôr em risco o modelo que está a ser construído. Para além disso, está associada a problemas como a perda de poder e precisão ao estar a formar novos grupos dentro de cada variável, isto é, ao não se ter em consideração a variação intra categórica reduzir-se-á o poder estatístico de um estudo (Greenland, 1995). Para além da perda de precisão de valores estimados e de probabilidades, a categorização assume que a relação entre a variável independente e a variável resposta é igual para os valores dentro de cada intervalo, o que na realidade pode não se verificar.

.

Alternativas Metodológicas

Splines Cúbicas

Uma abordagem que é muito comum na modelação da não linearidade das variáveis é dividir a variável contínua em categorias. Embora a categorização proporcione conclusões do modelo que possam ser mais atrativas, pode não descrever os dados da melhor maneira pois o poder preditivo diminui. O uso de pontos de corte não permite uma boa relação entre a variável explicativa e a variável resposta. À medida que o número de categorias aumenta, também o número de graus de liberdade aumenta o que é preocupante para a predição de um modelo. Ao invés de categorizar uma variável, como alternativa, inclui-se a variável explicativa como uma variável contínua, encontrando uma transformação que produza uma relação linear. Outra alternativa é usar um polinómio quadrático ou cúbico para modelar o relacionamento entre as variáveis (Croxford, 2016). Segundo Harrel (2017), uma melhor abordagem à discretização, que maximiza o poder preditivo e assume uma relação simples entre as variáveis independentes e a variável explicativa, é o uso de uma Função Spline Cúbica ou Função Spline de Regressão. Splines são funções formadas por diferentes polinómios de grau menor ou igual a um *m*, definidos para cada intervalo entre os pontos de interpolação de modo que em cada ponto de interpolação a spline seja contínua, assim como todas as derivadas até à ordem *m-1*.

Esta é então uma abordagem alternativa que divide o intervalo de aproximação em subintervalos e constrói um polinómio de aproximação, geralmente, diferente em cada subintervalo, ao qual se dá o nome de aproximação polinomial por partes. Nas situações em que o número de pontos de interpolação é grande, a inexatidão na aproximação obtida com um polinómio de grau elevado é denominada por erros de arredondamento. Quando a função que se quer interpolar possui derivadas de valor numérico elevado em alguma região do intervalo de interpolação, a aproximação é também prejudicada em todo o intervalo.

Na prática, normalmente são usadas Splines Cúbicas (polinómio de grau 3) por proporcionarem e permitirem uma flexibilidade suficiente ao ajuste dos dados, não exigindo tantos graus de liberdade como splines de grau superior. Uma vez que splines lineares apresentam descontinuidades nos nós da primeira derivada e splines quadráticas têm apenas a primeira derivada contínua (ou seja, a curvatura da spline pode ficar sujeita a uma troca nos nós), as splines cúbicas são usadas mais frequentemente (Ruggiero & Lopes, 1988). De acordo com Burden & Faires (2010), geralmente, um polinómio cúbico é constituído por quatro constantes, portanto há flexibilidade suficiente para assegurar que a interpolação não é apenas diferenciável no intervalo mas que também tem uma segunda derivada contínua.

Dada uma função f definida no intervalo $[a, b]$ e dado o conjunto de nós $a = x_0 < x_1 < \cdots < x_n = b$, uma Spline Cúbica S para f é uma função que satisfaz as seguintes condições:

- **a)** $S(x)$ é um polinómio cúbico, designado por $S_j(x)$, no subintervalo $\left[x_j, x_{j+1}\right]$ tal que $j = 0, 1, \dots, n-1$;
- **b)** $S_j(x_j) = f(x_j)$ e $S_j(x_{j+1}) = f(x_{j+1})$ para cada $j = 0, 1, \dots, n-1$;
- **c)** $S_{j+1}(x_{j+1}) = S_j(x_{j+1})$ para cada $j = 0, 1, \dots, n-2$;
- **d)** $S'_{j+1}(x_{j+1}) = S'_j(x_{j+1})$ para cada $j = 0, 1, \dots, n-2$;
- **e)** $S''_{j+1}(x_{j+1}) = S''_j(x_{j+1})$ para cada $j = 0, 1, \dots, n-2$;
- **f)** Um dos seguintes conjuntos de condições é satisfeito:
  - **i.** $S''(x_0) = S''(x_n) = 0$
  - **ii.** $S'(x_0) = f'(x_0)$ e $S'(x_n) = f'(x_n)$

Construção de uma Spline Cúbica

A spline definida num intervalo que é dividido em n subintervalos requer o cálculo de $4n$ constantes. Para se construir uma spline cúbica de uma função f as condições mencionadas anteriormente são aplicadas aos polinómios cúbicos

$$S_j(x) = a_j + b_j(x - x_j) + c_j(x - x_j)^2 + d_j(x - x_j)^3$$

para cada $j = 0, 1, \dots, n-1$. Desde que $S_j(x_j) = a_j = f(x_j)$ a condição **c)** pode ser aplicada para se conseguir obter

$$a_{j+1} = S_{j+1}(x_{j+1}) = S_j(x_{j+1}) = a_j + b_j(x_{j+1} - x_j) + c_j(x - x_j)^2 + d_j(x - x_j)^3$$

para cada $j = 0, 1, \dots, n-2$.

Considere-se que

$$h_j = x_{j+1} - x_j$$

para cada $j = 0, 1, \dots, n-1$. Se também for definido que $a_n = f(x_n)$ vem que

$$a_{j+1} = a_j + b_j h_j + c_j h_j^2 + d_j h_j^3$$

para cada $j = 0, 1, \dots, n-1$.

Da mesma maneira, considere-se que $b_n = S'(x_n)$ e observe-se que

$$S_j'(x) = b_j + 2c_j(x - x_j) + 3d_j(x - x_j)^2$$

implica que $S_j'(x_j) = b_j$ para cada $j = 0, 1, \dots, n-1$. Aplicando a condição **d)** resulta

$$b_{j+1} = b_j + 2c_j h_j + 3d_j h_j^2$$

para cada $j = 0, 1, \dots, n-1$.

Outra relação que permite o cálculo dos coeficientes de S_j é definir $c_n = \frac{S''(x_n)}{2}$ e aplicar a condição **e)**. Assim, para cada $j = 0, 1, \dots, n-1$ vem que

$$c_{j+1} = c_j + 3d_j h_j$$

Resolvendo a equação anterior em ordem a d_j e substituindo o seu valor nas equações que dizem respeito ao cálculo de a_{j+1} e de b_{j+1} resultam as novas equações definidas da seguinte forma:

$$a_{j+1} = a_j + b_j h_j + \frac{h_j^2}{3}(2c_j + c_{j+1})$$

$$b_{j+1} = b_j + h_j(c_j + c_{j+1})$$

para cada $j = 0, 1, \dots, n-1$.

Resolvendo a última equação que define a_{j+1} em ordem a b_j vem que

$$b_j = \frac{1}{h_j}(a_{j+1} - a_j) - \frac{h_j}{3}(2c_j + c_{j+1})$$

e consequentemente, depois de uma redução do índice para b_{j-1} resulta que

$$b_{j-1} = \frac{1}{h_j - 1}(a_j - a_{j-1}) - \frac{h_{j-1}}{3}(2c_{j-1} + c_j)$$

Substituindo estes valores na última equação que define b_{j+1}, com uma redução do índice, vem o sistema de equações linear definido por

$$h_{j-1}c_{j-1} + 2(h_{j-1} + h_j)c_j + h_j c_{j+1} = \frac{3}{h_j}(a_{j+1} - a_j) - \frac{3}{h_{j-1}}(a_j - a_{j-1})$$

para cada $j = 1, 2, \dots, n-1$.

Este sistema envolve apenas o cálculo de $\{c_j\}_{j=0}^{n}$. Os valores de $\{h_j\}_{j=0}^{n-1}$ e $\{a_j\}_{j=0}^{n}$ são dados, respetivamente, pelo espaço entre os nós $\{x_j\}_{j=0}^{n}$ e pelos valores de f nos seus nós. Após terem sido calculados os valores de $\{c_j\}_{j=0}^{n}$ basta determinar as constantes $\{b_j\}_{j=0}^{n-1}$ e $\{d_j\}_{j=0}^{n-1}$.

Assim é então possível construir os polinómios cúbicos.

A razão pela qual não se optou pela técnica descrita é o consumo de tempo que exigiria por ser necessária uma análise de variável em variável. Tal não compensaria comparativamente com a metodologia adotada.

Modelo Alternativo

Com o objetivo de comparar os resultados de um modelo com as variáveis no seu estado original (variáveis contínuas) com os resultados de um modelo com as mesmas variáveis, mas em que apenas uma das variáveis contínuas é discretizada, realizou-se um novo modelo logístico. Desta forma consegue-se estudar o impacto da discretização e concluir de que forma esta poderá influenciar os resultados finais.

O estudo deste novo modelo tem como objetivo, através de uma regressão logística, estimar a probabilidade de um Cliente fazer um determinado tipo de depósito a prazo. Assim a variável independente Y é igual a 1 se o cliente faz o depósito ou é igual a 0 se não faz. Os registos da base de dados provêm de Clientes de uma Instituição Bancária Portuguesa que foram contactados (por via telefone ou telemóvel) no sentido de lhes ser proposto esse mesmo depósito a prazo. Consoante os dados recolhidos, 11% dos Clientes dizem fazer o depósito a prazo e 89% não fazem.

Gráfico 7: Distribuição dos Clientes relativamente à proposta do depósito a prazo

O conjunto inicial de variáveis que podem influenciar ou não a decisão do Cliente aderir ao produto é constituído por:

1. *age*: idade do Cliente
2. *duration*: duração do último contacto (em segundos)
3. *campaign*: número de contactos realizados durante esta campanha para o Cliente
4. *previous*: número de contactos realizados antes desta campanha para o Cliente
5. *pdays*: número de dias que passaram depois do Cliente ter sido contactado para a última campanha
6. *nr.employed*: número de empregados
7. *cons.conf.idx*: índice de confiança do consumidor (indicador mensal)
8. *emp.var.rate*: taxa de variação de emprego (indicador trimestral)
9. *cons.price.idx*: índice de preços do consumidor (indicador mensal)
10. *euribor3m*: taxa Euribor de 3 meses (indicador diário)

Resultados do Modelo

Através do método Stepwise e calculando as estatísticas de Wald para determinarmos a significância de cada uma das variáveis chegamos aos resultados finais de cada um dos modelos.

Os resultados obtidos foram os seguintes:

Modelo Contínuo

Variável	$\widehat{\beta}$	$\widehat{SE}_{\widehat{\beta}}$	Estatística de Wald	p-value	$e^{\widehat{\beta}}$
Intercept	-115,6755	13,5040	-8,5660	<0,0001	5,79102E-51
duration	0,0049	0,0002	20,4691	<0,0001	1,0049
pdays	-0,0016	0,0002	-7,6840	<0,0001	0,9984
cons.conf.idx	0,0582	0,0114	5,1184	<0,0001	1,0600
emp.var.rate	-0,9840	0,0630	-15,6070	<0,0001	0,3738
cons.price.idx	1,2321	0,1443	8,5383	<0,0001	3,4285

Tabela 19: Variáveis selecionadas no modelo final contínuo

Através da tabela 19 pode-se concluir que as variáveis finais são as variáveis "duration", "pdays", "cons.conf.idx", "emp.var.rate" e "cons.price.idx". e que todas elas são significativas.

Se for calculada a probabilidade de um Cliente fazer o depósito a prazo considerando todas as variáveis independentes iguais a zero, isto é:

$$\frac{e^{\beta_0}}{1+e^{\beta_0}} = \frac{e^{-115.6755}}{1+e^{-115.6755}} = 5.791E-51 \simeq 0\%$$

pode ser concluído que excluindo quaisquer características associadas ao modelo o Cliente não faria o determinado depósito a prazo, ou seja, é o conjunto de características finais que influencia a adesão do Cliente ao depósito a prazo.

A matriz de confusão compara a classificação dos dados observados dos Clientes com os dados previstos pelo modelo. Tal como pode ser observado na tabela 20, de um total de 280 Clientes previstos como terem aderido ao depósito a prazo 92 Clientes (33%) foram previstos como positivos quando na realidade não o afirmam ter feito.

Matriz de Confusão				
		Observados		
		Positivos	Negativos	Total
Previstos	Positivos	188	92	280
	Negativos	263	3576	3839
	Total	451	3668	4119

Tabela 20: Matriz de Confusão do modelo contínuo

Relativamente ao Erro de Tipo I, ou seja, a probabilidade de um Cliente ter sido previsto como ter feito o depósito a prazo quando na realidade diz não o ter feito, este é de 2,51% tal como pode ser consultado na tabela 21.

Tipo I	2,51%
Tipo II	58,31%
Sensibilidade	41,69%
Especificidade	97,49%

Tabela 21: Erros, Sensibilidade e Especificidade do modelo final contínuo

Os extremos dos intervalos de confiança dos coeficientes estimados com um grau de confiança de 95% são os indicados na tabela 22.

	$\widehat{\beta_0}$	$\widehat{\beta_1}$	$\widehat{\beta_2}$	$\widehat{\beta_3}$	$\widehat{\beta_4}$	$\widehat{\beta_5}$
LS	-89,2082	0,0054	-0,0012	0,0805	-0,8604	1,5150
LI	-142,1428	0,0045	-0,0020	0,0359	-1,1075	0,9493

Tabela 22: Intervalos de Confiança dos coeficientes estimados

Para os intervalos de confiança do logit e das probabilidades logísticas, como são envolvidas as respetivas observações para serem calculados, optou-se por concretizar através de um exemplo.

Concretizando para a observação 40 vem que:

$$\widehat{VAR}[\hat{g}(40)] = x_{40}{}'(X'VX)^{-1}x_{40} = 0.0445$$

$$logit_{40} = g(40) = -1.8643$$

$$\hat{\pi}(40) = \frac{e^{logit_{40}}}{1 + e^{logit_{40}}} = 0.1342$$

Assim o intervalo de confiança para o logit ($\hat{g}(40)$) é:

$$(-1.8643 \pm 1.96 * 0.0445) = (-2.2777, -1.4510)$$

Consequentemente vem o intervalo de confiança para a probabilidade logística associada:

$$\left(\frac{e^{-2.2777}}{1 + e^{-2.2777}}, \frac{e^{-1.4510}}{1 + e^{-14510}}\right) = (0.093, 0.1898)$$

As amplitudes dos intervalos de confiança que são superiores a 10% manifestam-se em 486 indivíduos, ou seja, em 11.8% do total dos Clientes. Já para amplitudes superiores a 20% são encontrados um menor número de indivíduos, 34 Clientes. Logo as amplitudes dos intervalos de confiança não são muito grandes o que faz com que seja uma boa previsão face à realidade.

Num bom modelo a curva ROC deve crescer rapidamente para 1, à medida que o ponto de corte cresce, afastando-se da diagonal que corresponde a um modelo em que a previsão é feita completamente ao acaso. Assim esta curva representa para cada valor do ponto de corte a percentagem de verdadeiros positivos contra a percentagem de falsos positivos. Tal como pode ser visto no gráfico 8, este modelo apresenta uma curva ROC muito boa permitindo concluir que o modelo faz uma boa previsão dos Clientes que fazem ou não o depósito a prazo.

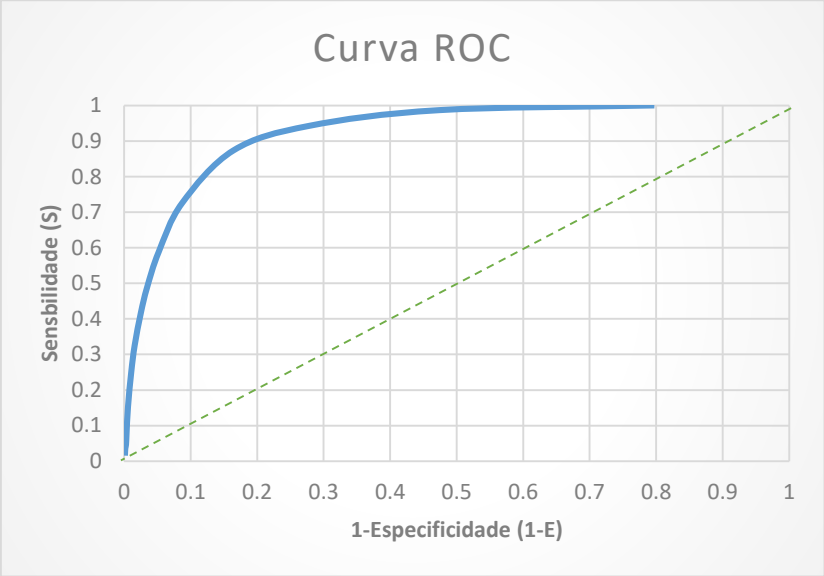

Gráfico 8: Curva ROC do modelo final contínuo

Modelo com a variável "duration" discretizada" e todas as outras variáveis independentes contínuas

A variável "duration" foi discretizada, consoante o tempo da chamada, nos seguintes intervalos:

- d_{80}: até 80 segundos
- d_{160}: de 80 a 160 segundos
- d_{240}: de 160 a 240 segundos
- d_{400}: de 240 a 400 segundos
- d_{1200}: de 400 a 1200 segundos

As variáveis finais deste modelo são as mesmas que do modelo anterior com exceção de uma categoria da variável "duration", ou seja, foram desprezados os dados da variável "duration" correspondentes aos valores registados numa chamada de 240 a 400 segundos. Todas as variáveis são significativas, como se pode ver na tabela 23.

Variável	$\hat{\beta}$	$\widehat{SE}_{\hat{\beta}}$	Estatística de Wald	p-value	$e^{\hat{\beta}}$
Intercept	-99,3500	13,7959	-7,2014	<0,0001	7,1257E-44
d80	-3,2595	0,5266	-6,1900	<0,0001	0,0384
d160	-1,8422	0,2229	-8,2633	<0,0001	0,1585
d240	-0,9931	0,1925	-5,1592	<0,0001	0,3704
d1200	1,5225	0,1524	9,9919	<0,0001	4,5839
pdays	-0,0017	0,0002	-7,5026	<0,0001	0,9983
cons.conf.idx	0,0640	0,0114	5,5958	<0,0001	1,0661

emp.var.rate	-0,8748	0,0614	-14,2546	<0,0001	0,4170
cons.price.idx	1,0830	0,1473	7,3498	<0,0001	2,9534

Tabela 23: variáveis selecionadas no modelo final contínuo com a variável duration discretizada

Se for calculada a probabilidade de um Cliente fazer o depósito a prazo considerando todas as variáveis independentes iguais a zero, isto é:

$$\frac{e^{\beta_0}}{1+e^{\beta_0}} = \frac{e^{-99.3500}}{1+e^{-99.3500}} = 7.1260E-44 \simeq 0\%$$

pode ser concluído que o conjunto de características finais influencia a adesão do Cliente ao depósito a prazo.

Tal como pode ser observado na tabela 24, de um total de 2059 Clientes previstos como tendo aderido ao depósito a prazo 1840 Clientes (89%) foram previstos como positivos quando na realidade não o afirmam ter feito. Em relação ao modelo anterior este resultado não retrata uma boa previsão do modelo.

Matriz de Confusão				
		Observados		
		Positivos	**Negativos**	**Total**
Previstos	**Positivos**	219	1840	2059
	Negativos	232	1828	2060
	Total	451	3668	4119

Tabela 24: Matriz de confusão do modelo final contínuo com a variável duration discretizada

Relativamente ao Erro de Tipo I, ou seja, a probabilidade de um Cliente ter sido previsto como ter feito o depósito a prazo quando na realidade diz não o ter feito, é de 50,16% tal como pode ser consultado na tabela 25.

Tipo I	50,16%
Tipo II	51,44%
Sensibilidade	48,56%
Especificidade	49,84%

Tabela 25: Erros, Sensibilidade e Especificidade modelo final contínuo coma variável duration discretizada

Os extremos dos intervalos de confiança dos coeficientes estimados com um grau de confiança de 95% são os indicados na tabela 26.

	$\widehat{\beta_0}$	$\widehat{\beta_1}$	$\widehat{\beta_2}$	$\widehat{\beta_3}$	$\widehat{\beta_4}$	$\widehat{\beta_5}$	$\widehat{\beta_6}$	$\widehat{\beta_7}$	$\widehat{\beta_8}$
LS	-72,3105	1,8212	-0,0013	-2,2275	-1,4052	-0,6158	0,0864	-0,7545	1,3718
LI	-126,3896	1,2239	-0,0021	-4,2916	-2,2791	-1,3704	0,0416	-0,9951	0,7942

Tabela 26: Intervalos de Confiança dos coeficientes estimados com a variável duration discretizada

Para os intervalos de confiança do logit e das probabilidades logísticas, como são envolvidas as respetivas observações para serem calculados, optou-se por concretizar através de um exemplo tal como no modelo anterior e os resultados obtidos foram:

$$I.C.[\hat{g}(40)] = (-4.25, -3.22)$$

$$IC\ [\hat{\pi}(40)] = (0.0385, 0.014)$$

As amplitudes dos intervalos de confiança que são superiores a 10% manifestam-se em 531 indivíduos, ou seja, em 12.9% do total dos Clientes. Já para amplitudes superiores a 20% são encontrados um menor número de indivíduos, 87 Clientes. As amplitudes dos intervalos de confiança, quando comparadas com o modelo anterior, pode-se concluir que são maiores.

Tal como pode ser visto no gráfico 9, este modelo apresenta uma curva ROC irregular e pior do que a curva ROC do modelo anterior, permitindo concluir que o modelo contínuo faz uma melhor previsão dos Clientes que fazem ou não o depósito a prazo relativamente ao modelo que tem a variável duration discretizada.

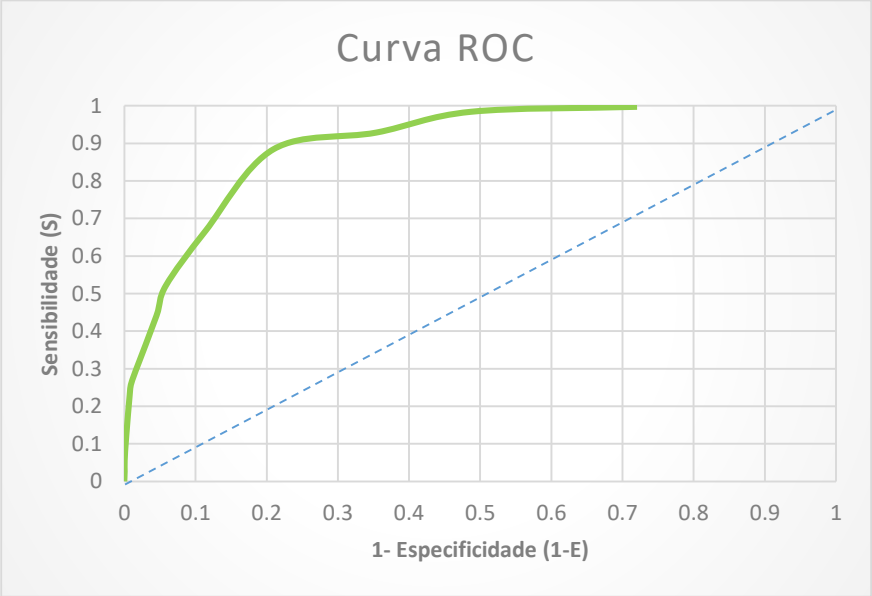

Gráfico 9: Curva ROC do modelo final continuo com a variável duration discretizada

Conclusão

O objetivo deste estudo é, em função do perfil de cada Cliente, analisar a forma como a Vinculação se expressa na sua relação do Cliente com o Banco, isto é, caracterizar a posse e utilização dos principais produtos financeiros de cada Cliente que levam a concluir acerca da vinculação do Cliente ao Banco. Com este propósito criou-se um modelo que pudesse prever o nível de vinculação de um Cliente do Banco B.

Inicialmente procedeu-se à identificação das variáveis seguindo-se o seu tratamento estatístico de forma a adequá-las à aplicação da modelação estatística. O processo de seleção de variáveis começou por uma análise das correlações, primeiro entre cada uma das variáveis, para que o modelo não tenha mais do que uma variável a conter exatamente a mesma informação, e em segundo entre cada uma das variáveis e a variável explicativa, de maneira a serem selecionadas as que tiverem maior correlação pois são essas que melhor explicam a variável dependente. Depois de ter sido feita esta primeira análise, através de técnicas como Árvores de Decisão, Qui-quadrado e R-quadrado foi feita uma segunda seleção de variáveis para se poder passar à construção do modelo. Todo este processo de avaliação das variáveis foi feito depois de todas as variáveis terem sido discretizadas, isto é, as variáveis contínuas passaram a discretas. Assim, foi facilitada a interpretação e conclusão das análises de cada variável, não esquecendo que esta é uma técnica essencial para o pré-processamento dos dados em termos de machine learning.

Aquando da construção do modelo, o último método aplicado para selecionar o subconjunto de variáveis significativas foi o método de seleção de variáveis Stepwise que é baseado num algoritmo estatístico que verifica iterativamente a importância de cada uma das variáveis através da significância estatística do seu coeficiente estimado. No caso de uma variável que seja importante para o modelo, mas que ainda não seja significativa, sendo possível categorizá-la, de maneira a estudar a importância de cada categoria, são feitos agrupamentos de categorias de forma a se conseguir obter um modelo significativo. Quando todos os coeficientes são considerados significativos e se chegou ao conjunto de variáveis finais interpreta-se o modelo que foi então construído.

Finalmente procede-se à avaliação do modelo construído através da Matriz de Confusão que representa as instâncias reais e as previstas pelo modelo, isto é, os falsos positivos, falsos negativos, verdadeiros positivos e verdadeiros negativos. Através destes valores ainda é possível calculas certas probabilidades que avaliam o modelo previsto, tal como a sensibilidade, especificidade e o cálculo dos erros tipo I e II. Não esquecendo o último método de avaliação do modelo, a Curva de Lift que permite avaliar o desempenho do modelo ao representar o rácio entre a proporção de Clientes vinculados e a resposta da variável dependente.

Através da Regressão Logística aplicada, este modelo tem cerca de 59% de capacidade de prever corretamente os Clientes não Vinculados e 84% de capacidade de prever corretamente os Clientes Vinculados.

Uma conclusão importante deste modelo, que foi aliás o seu principal objetivo, é que 65% dos Clientes que se dizem ser vinculados ao Banco foram previstos como sendo realmente vinculados. A probabilidade dos dados previstos estarem próximos dos reais é de 79%. Assim foi permitida a construção de um indicador com dois níveis de vinculação: o Cliente é Vinculado ou o Cliente não é Vinculado.

No desenvolvimento deste modelo algumas limitações que surgiram foi a discretização das variáveis. De facto, embora seja uma técnica que agiliza o desenvolvimento computacional, também altera no entanto a precisão e o poder estatístico do estudo. Os conhecimentos matemáticos sofrem uma certa "rutura epistemológica" por estarem a ser feitas mudanças na forma inicial de uma variável e o resultado final não é obtido de maneira tão exata quanto se as variáveis fossem trabalhadas conforme a sua verdadeira natureza. Como alternativa a este processo poderia ter sido adotada a técnica das Splines

Cúbicas. Esta técnica é uma aproximação que consiste em dividir uma função (descrita por uma variável) em subintervalos interpolando. Desta maneira consegue-se definir uma função para cada subintervalo e é feito um estudo mais preciso sobre cada variável. No entanto, por ter de ser feita uma análise de variável a variável, haveria um grande consumo de tempo comparativamente com a técnica de discretização das variáveis.

Não obstante foi feita a construção de um modelo alternativo que visa mostrar a diferença entre trabalhar com uma variável contínua e com uma variável discretizada. O objetivo do modelo alternativo vai ao encontro do modelo desenvolvido neste projeto: através de uma Regressão Logística e de um conjunto de dados disponibilizados, estimar a probabilidade de um Cliente fazer, ou não fazer, um determinado depósito a prazo. Depois de ter sido feita a seleção de variáveis através do método de seleção Stepwise, análise dos coeficientes estimados e posterior avaliação do modelo conclui-se que quando se trata de um modelo com variáveis continuas, a probabilidade de um Cliente fazer o depósito sabendo que na realidade não o fez é de 42% enquanto que na presença de pelo menos uma variável discretizada é de 49%. Através da análise da Curva ROC dos dois modelos pode-se verificar ainda que há uma melhor precisão preditiva quando não é feita a categorização das variáveis.

Em suma e acerca do modelo desenvolvido para este projeto, o indicador de Vinculação é constituído por 2 níveis. Futuramente, o modelo ficou de ser implementado e posto em prática. Posteriormente vai ser avaliada a proposta de um indicador que reflita 3 níveis de probabilidade de ser vinculado, ou seja, o Cliente apresenta uma vinculação alta, média ou baixa.

Referências bibliográficas

- ✓ Alpuim, T. (2014). *Apontamentos da Cadeira de Estatística*
- ✓ Alpuim, T. (2016). *Apontamentos da Cadeira de Modelos Lineares*
- ✓ Bação, F.L. (s.d.). *Data Mining. Pós-Graduação em Estudos de Mercado e CRM*. Lisboa: Universidade Nova de Lisboa, Instituto Superior de Estatística e Gestão de Informação
- ✓ Berry, L. (1983). *Emerging Perspectives on Services Marketing*. Chicago: America Management Association
- ✓ Breiman, L. (2001). *Random Forests*. EUA: University of California – Berkeley
- ✓ Breiman, L. (2002). *Manual on Setting Up, Using, and Understanding Random Forests V3.1*. University of California – Berkeley
- ✓ Burden, R. L. & Faires, J. D. (2010). *Numerical Analysis Ninth Edition*. Canada: Cengage Learning
- ✓ Camilo, C.O. & Silva, J.C. (2009). *Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas*. Instituto de Informática, Universidade Federal de Goiás
- ✓ Christie, P., Georges, J., Thompson, J. & Wells, C. (2015). *Applied Analytics Using SAS Enterprise Miner Course Notes*. USA: SAS Institute Inc.
- ✓ Coppock, D.D. (2002). *Why Lift? Data Modeling and Mining*. Review Online
- ✓ Cotidiano. (2013). *Econometria: Multicolinearidade*. Portugal: Portal da Educação Tecnologia Educacional Ltda.
- ✓ Cox, D. R. & Hinkley D.V. (1974). *Theoretical Statistic*s. London: Chapman and Hall
- ✓ Croxford, R. (2016). *Restricted Cubics Spline Regression: A Brief Introduction*. Toronto: Ontario. Institute for Clinical Evaluative Sciences
- ✓ Fayyad, U. M. et al. (1996). *Advances in knowledge discovery and data mining*. California: AAAI/The MIT
- ✓ Frank E. & Witten I. (1999). *Making Better Use of Global Discretization. Proceedings of the 16th International Conference on Machine Learning*. USA: Morgan Kaufmann Publishers Inc. San Francisco
- ✓ Gomes, B.M.V. (2011). *Previsão de Churn em Companhias de Seguros*. Universidade do Minho

- ✓ Gomes, J.J. (2015). *Apontamentos da Cadeira de Estatística Aplicada*

- ✓ Greenland S. (1995). *Avoiding power loss associated with categorization and ordinal scores in dose-response and trend analysis*. Los Angeles: Department of Epidemiology, UCLA School of Public Health

- ✓ Han, J. & Kamber, M. (2006). *The Data Mining: Concepts and Techniques 2nd ed*. San Francisco: Morgan Kaufmann Publishers

- ✓ Harrel, F. (2017). *Problems Caused by Categorizing Continuos Variables*. Acedido a 12 de Maio de 2017, em: http://biostat.mc.vanderbilt.edu/wiki/Main/CatContinuous

- ✓ Hosmer, D. W. & Lemeshow, S. (2000). *Applied Logistic Regression (2nd Edition ed.)*. New York: USA: A Wiley-Interscience Publication, John Wiley & Sons Inc.

- ✓ Lira, S.A. (2004). *Análise de Correlação: Abordagem Teórica e de Construção dos Coeficientes com Aplicações*. Paraná: Curitiba

- ✓ Liu H., Hussain F., Tan C. & Dash M. (2002). *Discretization: An Enabling Technique. Data Mining and Knowledge Discover*. Netherlands: Kluwer Academic Publishers

- ✓ Lopes, J.L. (2007). *Fundamental dos Estudos de Mercado – Teoria e Prática*. Edições Sílabo

- ✓ Mittal A. & Cheong L. (2002). *Employing Discrete Bayes Error Rate for Discretization and Feature Selection Tasks*. Proceedings of the 1st IEEE International Conference on Data Mining. Japan: IEEE

- ✓ Muhlenbach, F. & Rakotomalala, R. (2005). *Discretization of continuous attributes*. IGI Global

- ✓ Potts, W. J. E. & Patetta, M. J. (2000). *Predictive Modeling Using Logistic Regression Course Notes*. USA: SAS Institute Inc.

- ✓ Ruggiero, M.A.G. & Lopes, V.L.R. (1988). *Cálculo Numérico: Aspetos Teóricos Computacionais*. Pearson. Makron Books

- ✓ SAS, (s.d.) *Variable Selection Node*

- ✓ SAS, (s.d.). *Decision Tree Node*

- ✓ Schaefer, R.L. (1986). *Alternative estimators in logistic regression when the data are collinear*. USA: Journal of Statistical Computation and Simulation

- ✓ Sen, P.K. & Singer, J.M. (1993). *Large Sample Methods in Statistics. An Introduction with Applications*. New York: Chapman & Hall

- ✓ Setti, M. (2009). *O Processo de Discretização do Raciocínio Matemático na Tradução para o Raciocínio Computacional: Um Estudo de Caso no Ensino/Aprendizagem de Algoritmos*. Curitiba: Universidade Federal do Paraná

- ✓ Stigler, S.M. (2011). *Karl Pearson and the Rule of Three*. Department of University of Chicago

- ✓ Taniar, D. (2008). *Data Mining and Knowledge Discovery Technologies*. Wang, J., Hu, X. & Zhu, D. Minimizing the Minus Sides of Mining Data. IGI Global

- ✓ Taylor, J.R. (1997). *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements 2nd ed*. University Science Books

- ✓ *Tipos de Data Mining*. Acedido a 14 de Maio de 2017, em: http://paginas.fe.up.pt/~mgi99021/it/tipos.htm

- ✓ Turkman, M. A. A. & Silva G. L. (2000). *Modelos Lineares Generalizados – da teoria à prática*

- ✓ Universidade de Berkeley. (2011). *Econometrics Laboratory Software Archive. Regression Analysis*. California

- ✓ Vilares, M.J. & Coelho, P.S. (2011). *Satisfação e Lealdade do Cliente*. Escolar Editora

Anexos

Variáveis	Categorias	Prop. E[1's] - Prop. 1's Declarados
SD1	C1	0,0043pp
	C2	-0,0043pp
	C3	0,0000pp
SD2	C1	0,0000pp
	C2	0,0000pp
	C3	0,0104pp
	C4	0,0079pp
	C5	-0,0060pp
	C6	-0,0082pp
	C7	-0,0041pp
	C8	0,0000pp
SD3	C1	-0,0153pp
	C2	0,0301pp
	C3	-0,0146pp
	C4	-0,0002pp
SD4	C1	0,0034pp
	C2	-0,0052pp
	C3	0,0017pp
	C4	0,0001pp
SD5	C1	0,0104pp
	C2	-0,0028pp
	C3	-0,0013pp
	C4	-0,0057pp
	C5	-0,0001pp
	C6	-0,0005pp
	C7	0,0000pp
SE1	C1	0,0117pp
	C2	-0,0143pp
	C3	-0,0029pp
	C4	0,0068pp
	C5	0,0026pp
	C6	0,0049pp
	C7	-0,0051pp
	C8	-0,0024pp
	C9	-0,0013pp
SE2	C1	-0,0350pp
	C2	0,0145pp
	C3	0,0176pp
	C4	0,0061pp
	C5	0,0035pp
	C6	-0,0067pp
SE3	C1	-0,0011pp
	C2	0,0167pp
	C3	-0,0088pp
	C4	-0,0032pp
	C5	-0,0037pp
	C6	0,0000pp

	C7	0,0000pp
	C8	0,0000pp
	C9	0,0000pp
SE4	C1	0,0168pp
	C2	-0,0002pp
	C3	0,0016pp
	C4	0,0083pp
	C5	0,0131pp
	C6	0,0059pp
	C7	0,0001pp
	C8	-0,0012pp
	C9	-0,0444pp
COMP1	C1	-0,0205pp
	C2	0,0233pp
	C3	0,0078pp
	C4	-0,0023pp
	C5	0,0000pp
	C6	-0,0083pp
COMP2	C1	0,0356pp
	C2	-0,0015pp
	C3	0,0004pp
	C4	-0,0079pp
	C5	-0,0133pp
	C6	-0,0023pp
	C7	-0,0027pp
	C8	-0,0083pp
CT1	C1	-0,0037pp
	C2	0,0037pp
	C3	0,0065pp
	C4	-0,0065pp
CT2	C1	0,0018pp
	C2	0,0017pp
	C3	0,0029pp
	C4	0,0073pp
	C5	0,0036pp
	C6	0,0007pp
	C7	-0,0093pp
	C8	-0,0086pp
CT3	C1	0,0018pp
	C2	-0,0017pp
	C3	0,0077pp
	C4	0,0079pp
	C5	-0,0030pp
	C6	-0,0126pp
CT4	C1	-0,0224pp
	C2	0,0107pp
	C3	0,0043pp
	C4	0,0074pp
CT5	C1	0,0012pp
	C2	0,0010pp
	C3	0,0049pp
	C4	0,0086pp

C5	-0,0030pp
C6	-0,0126pp
CT6	0,0563pp
CT7	0,0377pp
CT8	0,0537pp
CT9	0,0055pp
CT10	0,0238pp
CT11	0,0129pp
CT12	-0,0102pp
CT13	-0,0050pp
CT14	-0,0019pp
CT15	-0,0067pp
CT16	-0,0008pp
CT17	-0,0029pp
CT18	-0,0005pp
CT19	0,0071pp
CT20	-0,0002pp
CT21	0,0107pp
CT22	0,0146pp

Legenda das Variáveis:
SD: Sóciodemográficas
SE: Sócioeconómicas
COMP: Comportamentais
CT: Carteira